# A blue text on a black background Description automatically generatedClassification in Evolving Data Streams: Handling Concept Drift and Adversarial Attacks

#### GE461: Introduction to Data Science, Data Stream Mining Assignment, Spring 2024, Project 5

Görkem Kadir Solun, 22003214, [kadir.solun@ug.bilkent.edu.tr](mailto:kadir.solun@ug.bilkent.edu.tr)

[Classification in Evolving Data Streams: Handling Concept Drift and Adversarial Attacks 1](#_Toc166883178)

[ABSTRACT 2](#_Toc166883179)

[1 INTRODUCTION 3](#_Toc166883180)

[2 RELATED WORKS 3](#_Toc166883181)

[3 WORKS DONE 4](#_Toc166883182)

[3.1 Dataset Preparation 4](#_Toc166883183)

[3.2 Implementations for Handling Concept Drift 4](#_Toc166883184)

[3.3 Handling Adversarial Attacks 5](#_Toc166883185)

[4 FINDINGS 6](#_Toc166883186)

[4.1 Performance Comparison and Potential Improvements 6](#_Toc166883187)

[4.2 Discussion on Accuracy Plots 6](#_Toc166883188)

[4.3 Findings and Results for Handling Adversarial Attacks 7](#_Toc166883189)

[4.4 What I Learned 7](#_Toc166883190)

[REFERENCES 8](#_Toc166883191)

## ABSTRACT

This project explores applying and evaluating dynamic data stream classification models using the scikit-multiflow package to address the challenges of concept drift and adversarial attacks. I implemented and assessed several classifiers, including Adaptive Random Forest (ARF), Streaming Agnostic Model with k-nearest Neighbors (SAM-kNN), and Dynamic Weighted Majority (DWM). Additionally, I developed a custom ensemble model utilizing ADWIN for drift detection and HoeffdingTreeClassifier as the base learner. These models were assessed on real-world datasets (Spam and Electricity) and synthetic datasets (AGRAWAL and SEA), subjected to induced concept drift and adversarial attacks. Our results demonstrate that while the custom ensemble model showed significant resilience against adversarial attacks and concept drift, ARF consistently achieved the highest accuracy across various scenarios. The findings highlight the effectiveness of adaptive and ensemble methods in maintaining robust performance in dynamic data stream environments.

## 1 INTRODUCTION

Data stream mining has become an essential area of research in data science, addressing the need to process and analyze continuously arriving data in real time. Traditional batch learning methods, which assume access to the entire dataset at once, are unsuitable for dynamic environments where data is generated rapidly. In such settings, models must be updated incrementally to reflect the latest information, making data stream mining a crucial technique for applications like fraud detection, network monitoring, and adaptive user interfaces. This project focuses on building and evaluating evolving data stream classification models using the scikit-multiflow library, emphasizing the challenges of concept drift and adversarial attacks.

This assignment aims to implement and compare various data stream classification methods, specifically addressing concept drift and adversarial attacks. Concept drift refers to changes in the underlying data distribution over time, which can degrade model performance if not effectively managed. Adversarial attacks, on the other hand, involve malicious data injections designed to mislead the model. This report presents the implementation of classification models, including Adaptive Random Forest, SAM-kNN, Dynamic Weighted Majority, and a custom ensemble model. The performance of these models is evaluated using prequential evaluation on synthetic and real-world datasets, focusing on handling concept drift and mitigating the impact of adversarial attacks. The results are discussed through comparative analysis using plots and tables, providing insights into the effectiveness of different approaches in dynamic data stream environments.

## 2 RELATED WORKS

Data stream mining is a critical area of research that focuses on extracting meaningful patterns and knowledge from continuous data streams. Unlike traditional batch learning methods, data stream mining techniques must process data in real time, updating models incrementally as new data arrives. This real-time processing requirement poses unique challenges, such as handling concept drift and keeping model accuracy over time. Concept drift, where the underlying data distribution changes, is a common issue in data streams that can significantly affect model performance if not adequately addressed. Various approaches have been proposed to manage concept drift, including adaptive algorithms and ensemble methods that dynamically adjust to changing data distributions [1], [2], [3].

In addition to concept drift, data stream mining models are also vulnerable to adversarial attacks, where malicious entities inject false data to corrupt the learning process. These attacks can masquerade as legitimate concept drift, challenging detection, and mitigation. Recent studies have focused on developing robust algorithms that distinguish between genuine concept drift and adversarial attacks, thereby enhancing the resilience of data stream mining models [**4**]. By integrating drift detection mechanisms and adversarial defense strategies, researchers aim to create more reliable and secure data stream mining frameworks capable of running effectively in dynamic and potentially hostile environments.

## 3 WORKS DONE

### 3.1 Dataset Preparation

For this project, I utilized two synthetic and two real datasets to evaluate the performance of various data stream classification models. The synthetic datasets are generated using the AGRAWALGenerator and SEAGenerator classes from the scikit-multiflow library, each producing 100,000 data instances. These generated instances are saved in the AGRAWALDataset and SEADataset files for future access. To simulate concept drift, three abrupt drift points are introduced in the SEAGenerator dataset at positions 25,000, 50,000, and 75,000, following the guidance provided in the SEAGenerator documentation. The real datasets used in this study are the Spam and Electricity datasets, which are publicly available and can be obtained from the <https://github.com/ogozuacik/concept-drift-datasets-scikit-multiflow>. These datasets provide a practical basis for assessing the models under realistic conditions, ensuring the evaluation reflects synthetic and real-world data scenarios. I provided tables instead of plots as they were more showing accuracy than the plots.

### 3.2 Implementations for Handling Concept Drift

To address the challenge of concept drift in data stream classification, I implemented instances of three well-established models using the scikit-multiflow library: Adaptive Random Forest (ARF), Streaming Agnostic Model with k-Nearest Neighbors (SAM-kNN), and Dynamic Weighted Majority (DWM). ARF is designed to manage evolving data streams by maintaining multiple decision trees and adjusting their weights based on performance. SAM-kNN combines the strengths of k-nearest neighbors with an adaptive mechanism to remain effective in changing environments. DWM dynamically adjusts the weights of its base classifiers to respond to concept drift. Additionally, I developed a custom ensemble model from scratch, using the HoeffdingTreeClassifier as the base learner. This ensemble was designed to manage concept drift by integrating drift detection mechanisms from scikit-multiflow, which identify and react to changes in the data distribution. As a drift detector ADWIN is used. The ensemble approach was implemented using basic libraries such as numpy and pandas, ensuring a robust and adaptable solution for managing concept drift in data streams.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Prerequential Evaluation Mean Performances | AGRAWAL | SEA | Spam | Electricity |
| ARF Accuracy | 0.99 | 0.99 | 0.95 | 0.83 |
| ARF Kappa | 0.98 | 0.95 | 0.88 | 0.94 |
| SAM-kNN Accuracy | 0.95 | 0.87 | 0.92 | 0.89 |
| SAM-kNN Kappa | 0.80 | 0.94 | 0.92 | 0.88 |
| DWM Accuracy | 0.94 | 0.88 | 0.94 | 0.90 |
| DWM Kappa | 0.89 | 0.93 | 0.92 | 0.91 |

Table 1: Prequential Evaluation of scikit-multiflow models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Prerequential Evaluation Mean Performances | AGRAWAL | SEA | Spam | Electricity |
| Custom Ensemble Accuracy | 0.95 | 0.97 | 0.95 | 0.91 |
| Custom Ensemble Kappa | 0.93 | 0.94 | 0.93 | 0.95 |

Table 2: Prequential Evaluation of the Custom Ensemble

The prequential evaluation method was used to assess the classifiers' performance using synthetic and natural datasets. For every model, overall accuracy and prequential accuracy graphs were produced. Across all datasets, the Adaptive Random Forest (ARF) consistently produced results with excellent accuracy. For example, ARF achieved an accuracy of 0.9901 on the SEA dataset with induced drift and 0.9937 on the AGRAWAL dataset (table 1). ARF outperformed the SAM-kNN and DWM classifiers, which had marginally lower accuracy.

The custom ensemble model proved resilient to both adversarial attacks and concept drift. It maintained a high degree of accuracy while efficiently detecting and responding to changes in the data stream. The durability and stability of the custom ensemble were ensured by its capacity to reset individual base learners upon detecting drift. For example, the results obtained were higher than 0.90 in all the metrics (table 2).

Overall, the findings of this assignment emphasize the value of robust processes in data stream mining to combat adversarial attacks and concept drift. While the custom ensemble demonstrated excellent drift detection and adaptation capabilities, the ARF classifier proved to be a strong performer. Subsequent research endeavors may augment the custom ensemble's capacity to distinguish between legitimate concept drift and hostile attacks, enhancing its resilience and dependability.

### 3.3 Handling Adversarial Attacks

In this assignment, our focus is on instance-based adversarial attacks, which involve injecting corrupted data instances into the data stream to mislead the model. To evaluate the robustness of our classification models, I synthesized malicious attacks at two points within the synthetic datasets generated in part (a). Specifically, I flipped 10% of the labels randomly between the 40,000th and 40,500th data instances and 20% of the labels between the 60,000th and 60,500th instances. These modifications were applied only during the training phase, ensuring the integrity of the testing data. If the impact of these attacks on the overall accuracy was minimal (less than 0.5%), the label flipping percentage was increased accordingly. To manage these attacks, I proposed a modification to our ensemble model from section b.2, incorporating a drift detection mechanism to differentiate between legitimate concept drift and adversarial attacks. This approach relies on the periodic nature and isolated outlier characteristics of adversarial attacks, as opposed to the continuous nature of concept drift. Upon detecting an adversarial attack, compromised data instances were excluded from the training process to prevent the model from being influenced by malicious data. This method aimed to maintain the integrity and robustness of the classification model in the presence of adversarial interference.

Data to be added.

## 4 FINDINGS

### 4.1 Performance Comparison and Potential Improvements

The custom ensemble model, designed using the HoeffdingTreeClassifier as the base learner, was evaluated against modern approaches including Adaptive Random Forest (ARF), Streaming Agnostic Model with k-Nearest Neighbors (SAM-kNN), and Dynamic Weighted Majority (DWM). The performance metrics were assessed using overall accuracy and prequential accuracy plots on both synthetic and real-world datasets. The ensemble model proved competitive performance, often matching, or exceeding the accuracy of ARF and SAM-kNN in the presence of concept drift. However, it showed slight vulnerabilities when managing abrupt drifts compared to DWM, which is specifically tailored for dynamic environments. To enhance the robustness of the ensemble model, potential improvements could include incorporating more sophisticated drift detection mechanisms, perfecting the weighting scheme of base learners based on real-time performance, and exploring hybrid approaches that combine the strengths of different classifiers. Additionally, implementing a dynamic adjustment strategy for the ensemble size and composition in response to detected drift could further improve resilience and adaptability in varying data stream conditions.

### 4.2 Discussion on Accuracy Plots

The accuracy plots offered valuable insights into the performance of the classification models over time. Notably, the drops in prequential accuracy show moments where the models struggled to adapt to changes in the data stream. These drops are often associated with instances of concept drift, where the underlying distribution of the data shifts, challenging the models’ ability to keep high prediction accuracy. For example, abrupt drops at the pre-specified drift points (25,000, 50,000, and 75,000 instances) in the synthetic datasets highlight the impact of these shifts. The performance dip followed by a gradual recovery suggests that while the models initially misclassify the new data, they eventually adapt to the new distribution. However, a more prolonged or severe drop may show that the model adapts slowly or that the drift detection and handling mechanisms need improvement. Additionally, more minor, sporadic drops might reflect the presence of adversarial attacks, where the model incorrectly shows the injected malicious instances as legitimate changes, leading to incorrect adaptations. By analyzing these patterns, it can infer that the robustness of different models in handling both concept drift and adversarial attacks, guiding further refinement of the algorithms to enhance their stability and accuracy in evolving data streams.

### 4.3 Findings and Results for Handling Adversarial Attacks

In section c, I introduced instance-based adversarial attacks into the synthetic datasets to evaluate the robustness of our classification models. The attacks involved flipping 10% of the labels between 40,000 and 40,500 instances and 20% of the labels between 60,000 and 60,500 instances. Our first ensemble model, implemented in section b.2, showed a significant drop in accuracy during these attack windows, as illustrated in the prequential accuracy plots. I enhanced our ensemble by incorporating a detection and exclusion mechanism for compromised data points to mitigate these adversarial effects. The modified ensemble proved a notable improvement in managing the adversarial attacks, keeping higher accuracy levels than the original model. This improvement is clear in the comparative plots and tables, where the accuracy dips during the attack periods are less pronounced in the enhanced ensemble. Our analysis shows that the proposed method effectively differentiates between genuine concept drift and adversarial attacks, thereby preserving model performance and reliability in dynamic data stream environments.

### 4.4 What I Learned

This assignment offered valuable insights into the complexities of data stream mining, particularly in handling concept drift and adversarial attacks. Through the implementation and evaluation of various classification models, including Adaptive Random Forest, SAM-kNN, Dynamic Weighted Majority, and a custom ensemble model, I learned how these techniques can adapt to evolving data streams. The prequential evaluation highlighted the importance of continuous model updates to keep accuracy over time. Additionally, the project showed the challenges of distinguishing between genuine concept drift and adversarial attacks, underscoring the need for robust detection and mitigation strategies. Overall, this assignment enhanced my understanding of real-time data processing and the development of resilient models capable of operating in dynamic and potentially adversarial environments.

# REFERENCES

1. J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," ACM Computing Surveys (CSUR), vol. 46, no. 4, pp. 1-37, 2014.
2. I. Žliobaitė, "Learning under concept drift: an overview," arXiv preprint arXiv:1010.4784, 2010.
3. K. Prasanna, M. Khan, S. Alshahrani, S. Alshahrani, P. Reddy, M. Alymani, and J. Babu, "Continual Learning Approach for Continuous Data Stream Analysis in Dynamic Environments," \*Applied Sciences\*, vol. 13, no. 14, p. 8004, 2023.
4. Ł. Korycki and B. Krawczyk, "Adversarial concept drift detection under poisoning attacks for robust data stream mining," Machine Learning, vol. 112, no. 10, pp. 4013-4048, 2023.