Mustafa Cankan Balcı

22101761

18 May 2025

GE461 Introduction to Data Science Project 5

Introduction

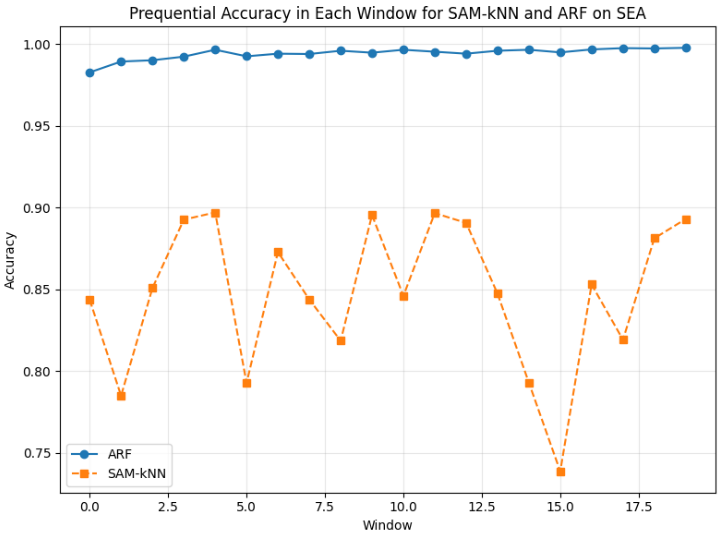
The traditional batch learning approaches are often inadequate for real-time application in the continuous data generation. For instance, user preferences change over the time, so the recommendation algorithm Data stream mining addresses this challenge by enabling incremental learning from continuously arriving data. This assignment aims to the design and evaluate stream-based classification models that can adapt to evolving data environments. Generated synthetic dataset generated and saved to a file for future use. There are two types of datasets used in the assignment: synthetic (AGRAWAL, SEA) and real world (Spam and Electricity). Generated synthetic dataset saved in the file for future access. In the assignment, there different types of approaches applied, which built-in functions (Adaptive Random Forest (ARF) and Streaming Agnostic Model with k-Nearest Neighbors (SAM-knn)) and ensemble classifier that integrates both active and passive methods for concept drift detection.

Dataset as Data streams

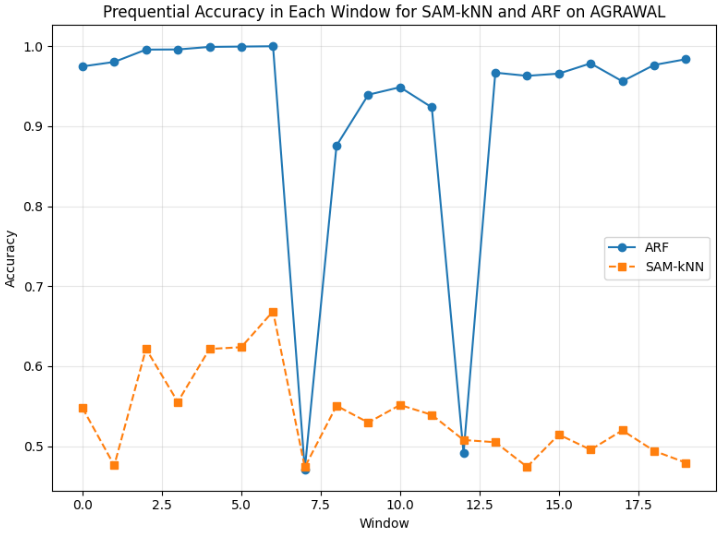
In this assignment, two synthetic and two real world datasets used to evaluate the data stream classification models. The synthetic datasets are generated using AGRAWAL Generator and SEA Generator from scikit-multiflow. Both datasets consist of 100,000 instances with two abrupt concept drifts at 35,000 and 60,000 positions. Also, the generated data saved for future use. For real-world evaluation, the Spam and Electricity datasets are employed.

Classification Task with Concept Drift Handling

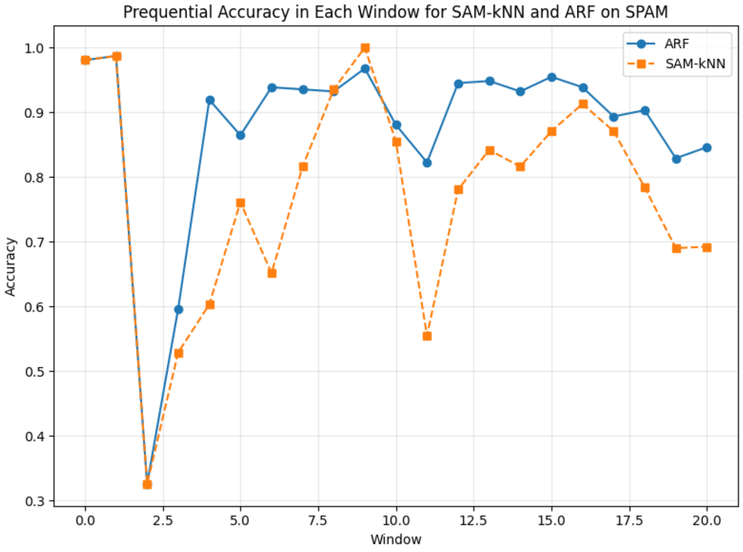
4.1) In this section, built-in classification models were utilized, and the test-then-train method was adopted for the training process. Initially, each model was pre-trained on a batch of 2,000 samples to establish awareness of the class labels. Subsequently, the models were updated using the test-then-train approach, where each incoming data chunk was first used for prediction and then for training. The data stream was divided into 20 equally sized windows. For each window, predictions and corresponding true labels were recorded to compute both window-level and overall classification accuracy. A noticeable drop in accuracy was observed at drift points, indicating the models' sensitivity to concept drift.



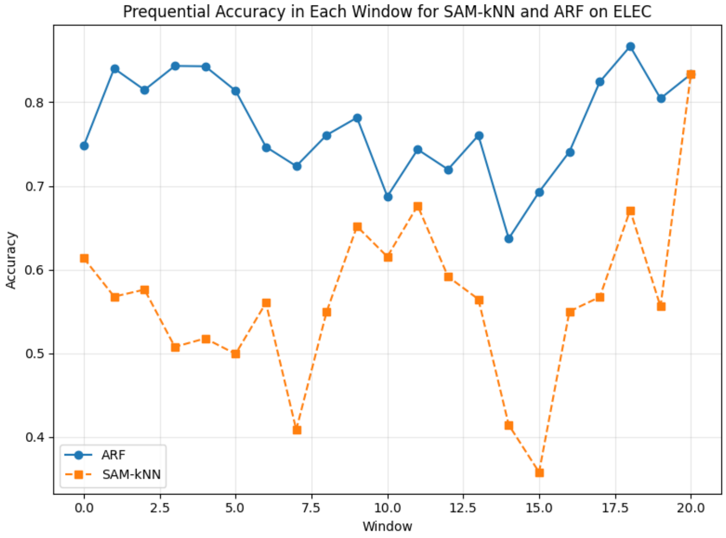
Graph 1: Prequential Accuracy in Each Window for SAM-knn and ARF on SEA (window size = 20)



Graph 2: Prequential Accuracy in Each Window for SAM-knn and ARF on AGRAWAL (window size = 20)

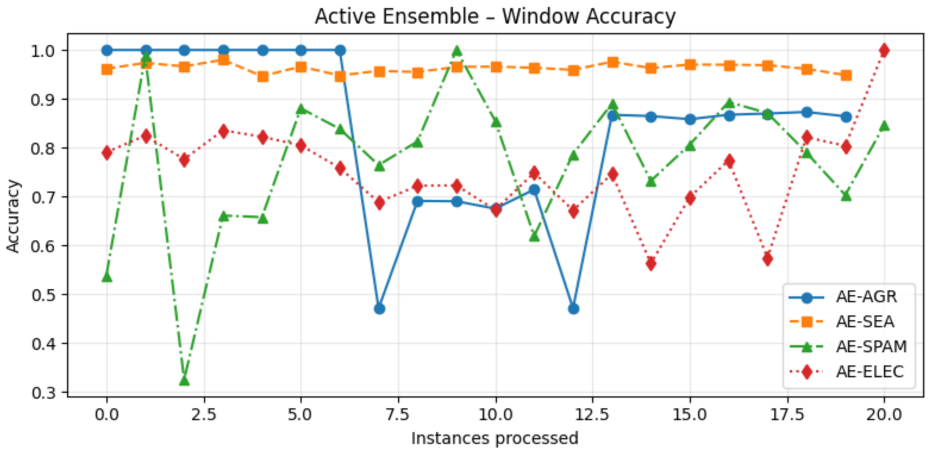


Graph 3: Prequential Accuracy in Each Window for SAM-knn and ARF on Spam (window size = 20)



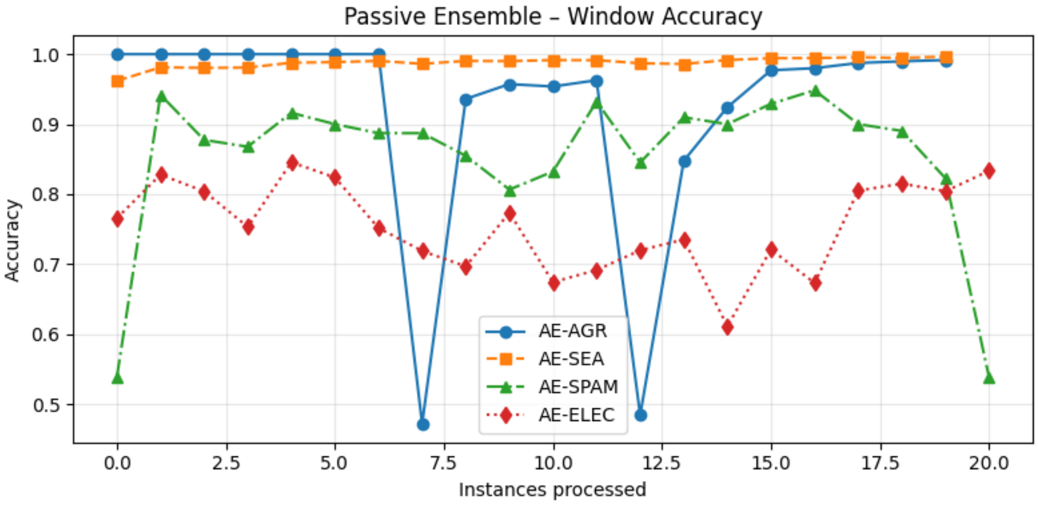
Graph 4: Prequential Accuracy in Each Window for SAM-knn and ARF on Electricity (window size = 20)

4.2) The ensemble classification model with active drift detection is used to handle concept drift in evolving data streams. Each base learner in the ensemble is associated with a distinct drift detector, which includes DDM, EDDM, and ADWIN. These detectors are linked to learners using a round-robin strategy to ensure diversity in drift detection. Each learner has a HoeffdingTreeClassifier that keeps track of recent prediction outcomes in a sliding window to ensure accuracy. Adaptive weighting is applied to each learner based on their performance across the sliding window, and the final ensemble prediction is determined by a weighted majority vote.



Graph 5: Prequential Accuracy in Each Window for
Active Ensemble with Each Dataset (window size = 20)

4.3) In this part of assignment, passive drift detection implemented to detect concept drift without depending on external drift detectors such as ADWIN. In the passive detection, model infer concept drift based on their performance history. Each learner keeps a sliding window of recent prediction results. If the window is full and accuracy falls below a predefine during weight majority voting. In the assignment, the threshold was 70 %. When passive drift detected, the underperforming learner was immediately replaced with a new HoeffdingTreeClassifier and retrained using the current data chunk to ensure smooth recovery.



Graph 6: Prequential Accuracy in Each Window for
Passive Ensemble with Each Dataset (window size = 20)

Results and Discussion

	SEA	AGRAWAL	Spam	Electricity
ARF	0.9888	0.9123	0.8710	0.7646
SAM-knn	0.8527	0.5375	0.7743	0.5644
Active Ensemble	0.9623	0.8390	0.7744	0.7536
Passive Ensemble	0.9733	0.9232	0.8250	0.7544

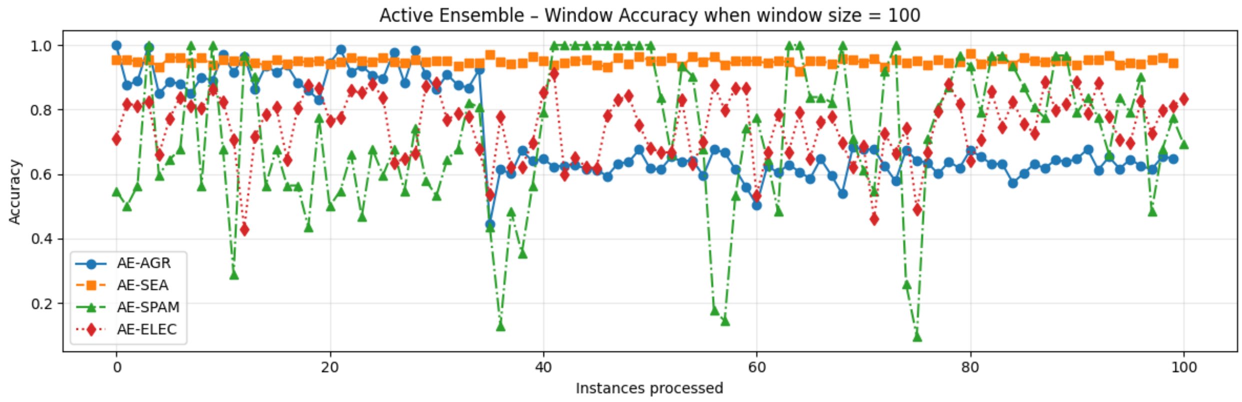
Table 1: Overall Accuracy for Each Stream Based Machine Learning Models (window size = 20)

5.1) ARF performed generally better than all of the ensemble models. However, SAM-knn did not perform as good as other models due to is limited capability to capture abrupt changes in data distribution. Thus, ensemble model performed achieved accuracy levels near to state-of-the art approaches. Improvements could include hybrid approaches which combined active and passive detection strategies, so less false positives detected by confidence-aware drift responses. The prequential accuracy plots (Graphs 1–6) further illustrate the dynamic response of each model to concept drift. In the drops of prequential accuracy plot, the concept drift occurs at these positions. Especially in 7 and 12 window these drops occurred since these window locations were drift positions in the dataset. In ensemble models, both active and passive variants showed dips in accuracy at these same windows, but the passive ensemble recovered more gracefully, with fewer false alarms triggered by noise.

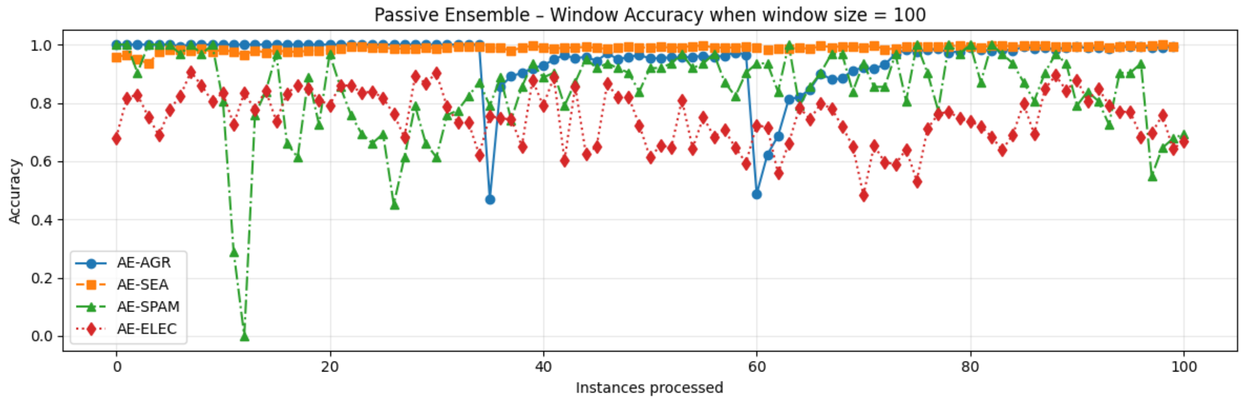
5.2) The different window sizes experimented in both the active and passive ensemble model to analyze their impact on the prediction accuracy. A smaller window size made the ensemble more susceptible to transient variations. As a result, even little changes were frequently mistaken for concept drift. In certain instances, this problem resulted in decreased overall accuracy and increased instability. Additionally, the ensemble model was more resilient to transient fluctuations and noise. However, the ensemble responded to real concept drifts more slowly as a result of this decreased sensitivity. Consequently, the model's ability to adjust over time failed, leading to a delayed recovery and decreased short-term accuracy. When table 1 and table 2 compared, these results could be obtained. Also, graph 7 and 8 showed larger window, and graph 5 and 6 reflects results in narrower window.

	SEA	AGRAWAL	Spam	Electricity
Active Ensemble	0.9499	0.7250	0.7276	0.7502
Passive Ensemble	0.9877	0.9540	0.8464	0.7492

Table 2: Overall Accuracy for Ensemble Models (window size = 100)



Graph 7: Prequential Accuracy in Each Window for Active Ensemble with Each Dataset (window size = 100)



Graph 8: Prequential Accuracy in Each Window for Passive Ensemble with Each Dataset (window size =100)

5.3) Passive Ensemble generally outperformed the Active Ensemble in terms of accuracy and stability according to table 1. Difference between ensemble approaches occurred due to reaction speed and reliability of drift detection. Passive Ensemble demonstrated better resistance to noise fewer false drift alarms. This approach provided greater resilience to short term fluctuations and produced fewer false drift alarms. On the other hand, Active Ensemble depended on external detectors (DDM, EDDM, ADWIN). These drift detectors were more prone to misclassifying temporary noise as a concept drift. As a result, this instability reduced overall accuracy in datasets with noisy or less structured patterns.

5.4) In the assignment, different types of approaches of stream-based machine learning models designed by using the real world and synthetic datasets. Table 1 demonstrates that, among the state-of-the-art models, the Adaptive Random Forest (ARF) continuously performed better than the SAM-kNN in terms of overall accuracy. Also, drift handling mechanisms were important for maintain the accuracy. Passive Ensemble model showed the better accuracy and stability than active ensemble. Additionally, hyperparameters such as window size crucially affected the model's performance and should be carefully adjusted based on the application. Overall, this project enhanced understanding of stream-based classification, concept drift and trade-offs involved in real time adaptive learning.