Margin Density Based Drift Detection

A comparative study

Baptiste Andre (B.G.L.Andre@student.tudelft.nl)
Supervisor: Lorena Poenaru-Olaru
Responsible Professor: Jan Rellermeyer
CSE 3000 9/12/2023

TU Delft

1.Introduction

Research Question:

How well do Margin Density (MD)-based concept drift detectors identify concept drift in case of synthetic/real-world data?

Concept drift: An incoming data distribution does not represent data distribution from training data within the context of deployed machine learning algorithms.

Why study Unsupervised Drift Detectors?

- Training Labels are Expensive therefore we study unsupervised drift detectors, which do not require labels on testing data.
- Better Marchine learning classification performance over time
- Insights about the data

Current literature:

- Mostly Descriptions of novel Drift Detectors
- Few Comparative analysis
- Some Overview Literature

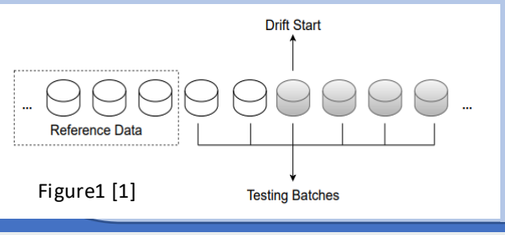
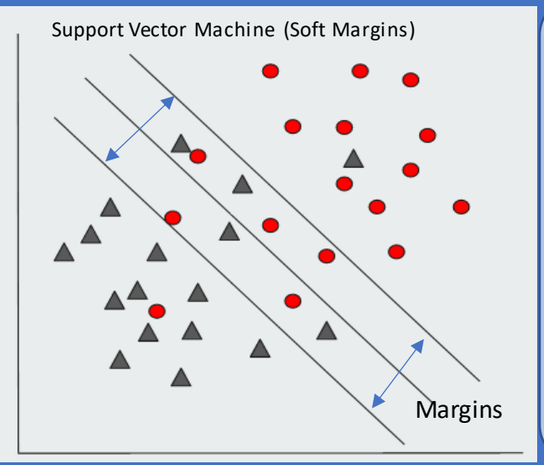


Figure1 [1]

Training is done on reference data, then detectors try to find drifts in the remaining data batches (or sliding window). Figure 1 represents the setup for the synthetic data used.



Margin Density (MD): data within margins/total data. For

Blindspot Density: The amount of data with classifying probabilities p(x)-p(y) <.5 divided by the total data

Fuzzy Margin Density: Same as Blindspot density, but when a data point is inside, it is scaled by factor .5(cos(p(x)-p(y)) + 1)

2.Methodology

Pre-processing:

- Min Max Scaler
- Feature encoders: Ordinal, Target, One – Hot
- For the Drift detectors: parameters recommended from the paper
- For real world data, we had to find the drifts

MD3_V1: [2]

1. Get MD of training data
2. Compare it to the batch MD
3. Keep track of max and min encountered MD
4. If max-min above threshold {parameter from paper}, drift detected

MD3_V2: [3]

1. 5-fold CV: Expected MD, Standard Deviation (SD), Accuracy
2. Is batch MD> than training MD + x(parameter from paper) * SD?
3. If yes, drift detected, check accuracy loss.

MD3_X: [3]

The algorithm replaces an SVM with an ensemble classifier and uses the prediction probabilities as a substitute metric for margin density (Blindspot density). The rest of the algorithm stays the same as with MD3_V2

Fuzzy Margin Density (FMD):[4]

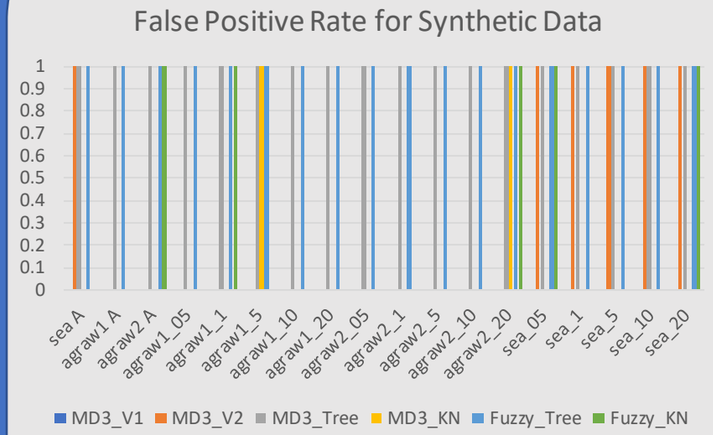
This algorithm replaces the margin density measure of MD3_X by utilizing fuzzy set theory. It replaces Blindspot density by FMD.

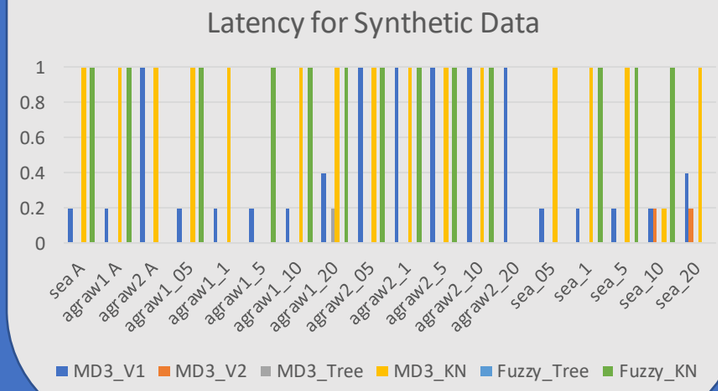
FPR: Measure of incorrectly classified drifts

Latency(Synthetic): Measure of how late a drift is detected after the first drifting batch.

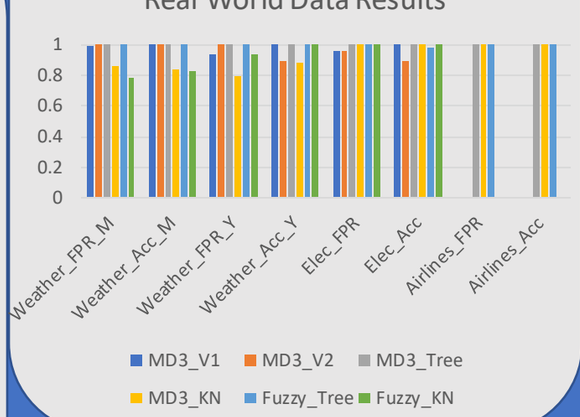
Accuracy (Acc): Total found drifting batches divided by total drifting batches.

3.1 Synthetic Data Results





3.2 Real World Data Results



4.Conclusion and Future Research

Main Findings:

- With correct parameter tuning, MD3_V1 and MD3_V2 can accurately detect drifts in the first batch in synthetic data. They might however not be well suited to find more than the first drift in an unsupervised environment.
- Ensemble detectors do not work on the used data sets. This could be due to its low dimensionality.

Future research:

- How do multi class margin density detectors perform?
- How should we tune parameters?
- Which Ensembles work best?
- How well do margin density detectors detect first drift in real world data?

References

https://github.com/bbonjean/MD3

[1] Poenaru-Olaru, L., Cruz, L., van Deursen, A., & Rellermeyer, J. S. (2022). *Are Concept Drift Detectors Reliable Alarming Systems? -- A Comparative Study*. doi:10.48550/ARXIV.2211.13098

[2] Sethi, T. S., & Kantardzic, M. (2015). Don't Pay for Validation: Detecting Drifts from Unlabeled data Using Margin Density. *Procedia Computer Science*, *53*, 103–112. doi:10.1016/j.procs.2015.07.284

[3] Sethi, T. S. and M. Kantardzic (2017). "On the reliable detection of concept drift from streaming unlabeled data." Expert Systems with Applications 82: 77-99.

[4] Jing Yang, Jie Zhang, and Sujuan Qin. A concept drift detection algorithm based on fuzzy marginal density, 2020.