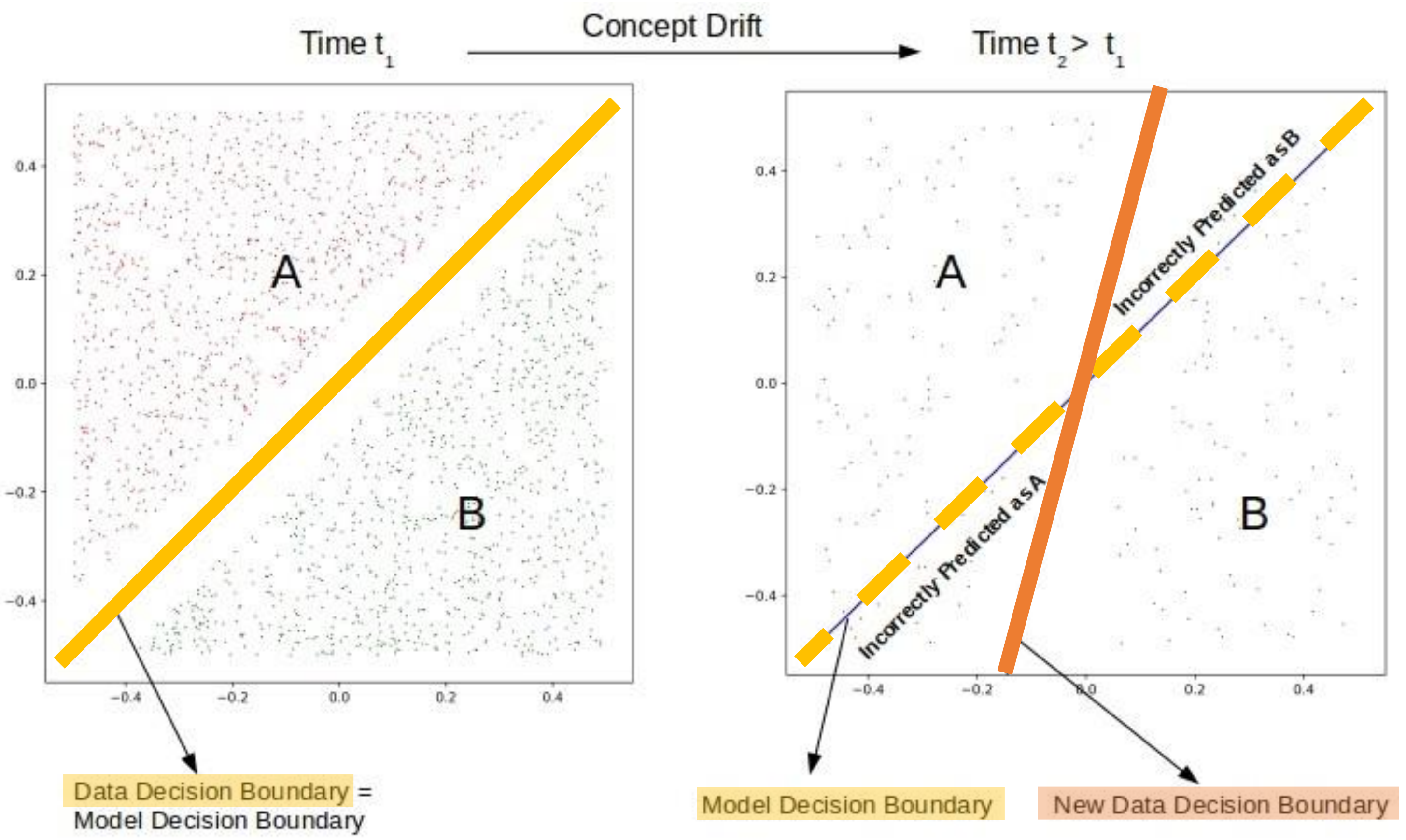
A Decision-Based Dynamic Ensemble Selection Method for Concept Drift

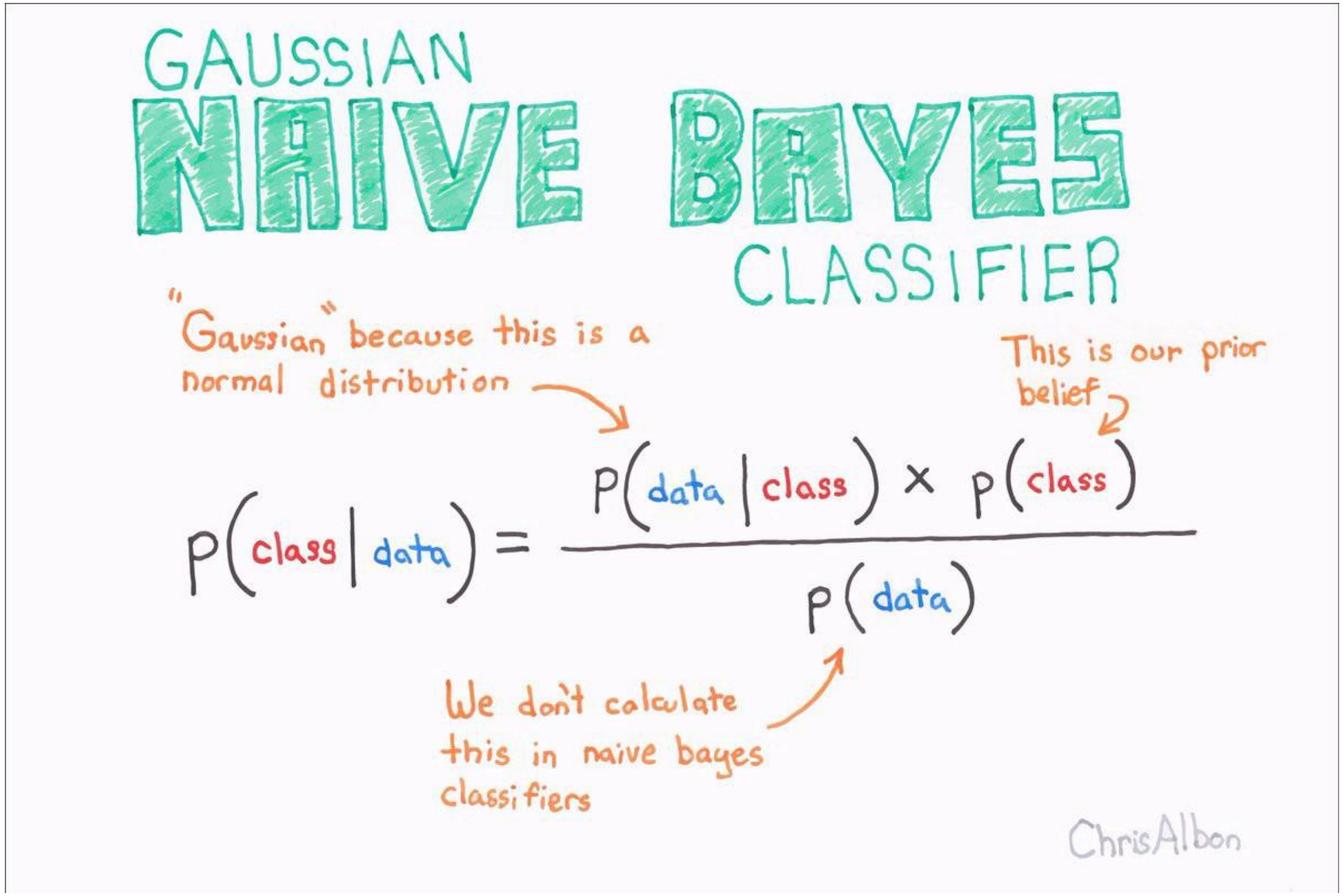
Tyng-Jiun Kuo, Yaser Alagele

Concept Drift



when data are continuously generated in streams, data and target concepts may change over time.

Gaussian Naive Bayes



- ➢ Used in a wide variety of classification tasks. Such as filtering spam, classifying documents, sentiment prediction etc.
- ➢ assumes the features that go into the model are independent of each other.

DDM(Drift Detection Method)

- Models the number of classification errors with a Binomial distribution.
- Each iteration an online classifier predicts the decision class of an example

DDM(Drift Detection Method)

Pi: probability of false prediction
Si: Standard deviation

Registers for tracking error rate

Check if the prediction is correct

$$s_i = \sqrt{\frac{p_i(1-pi)}{i}} \quad (1)$$

Input:

- S: a data stream of examples
- C: classifier

Output: W : a window with examples selected to train classifier C

DDM: Drift Detection Method

Initialize $(i, \boxed{p_i, s_i}, ps_{min}, \boxed{p_{min}, s_{min}})$;
$newDrift \leftarrow false$;
$W \leftarrow \emptyset$;
$W' \leftarrow \emptyset$;
for all examples $x_i \in S$ **do**
  if prediction $C(x_i)$ is incorrect **then**
    $p_i \leftarrow p_i + (1.0 - p_i)/i$;
  else
    $p_i \leftarrow p_i - (p_i)/i$;
  compute s_i using (1);
  $i \leftarrow i + 1$;
  if $i > 30$ (approximated normal distribution) **then**
    if $p_i + p_s \leq ps_{min}$ **then**
      $p_{min} \leftarrow p_i$;
      $s_{min} \leftarrow s_i$;
      $ps_{min} \leftarrow p_i + s_i$;
    if drift detected (3) **then**

Leibniz Universität Hannover

DDM(Drift Detection Method)

Warning Level Condition

$$p_i + s_i \geq p_{min} + \alpha s_{min} \quad (2)$$

Alarm Level Condition

$$p_i + s_i \geq p_{min} + \beta s_{min} \quad (3)$$

Update values for register is needed

Input:

- S: a data stream of examples
- C: classifier

Output: W : a window with examples selected to train classifier C

DDM: Drift Detection Method

Initialize $(i, p_i, s_i, ps_{min}, p_{min}, s_{min})$;
$newDrift \leftarrow false$;
$W \leftarrow \emptyset$;
$W' \leftarrow \emptyset$;
for all examples $x_i \in S$ **do**
  if prediction $C(x_i)$ is incorrect **then**
    $p_i \leftarrow p_i + (1.0 - p_i)/i$;
  else
    $p_i \leftarrow p_i - (p_i)/i$;
  compute s_i using (1);
  $i \leftarrow i + 1$;
  if $i > 30$ (approximated normal distribution) **then**
    if $p_i + p_s \leq ps_{min}$ **then**
      $p_{min} \leftarrow p_i$;
      $s_{min} \leftarrow s_i$;
      $ps_{min} \leftarrow p_i + s_i$;
    if drift detected (3) **then**

DDM(Drift Detection Method)

Alarm level reached

Warning level reached

```
Initalize (i, p_i, s_i, ps_min, p_min, s_min;
W ← W';
W' ← ∅;
else if warning level reached (2) then
  if newDrift = true then
    W' ← ∅;
    newDrift ← false;
  W' ← W' ∪ x_i;
else
  newDrift ← true;
W ← W ∪ x_i;
```

Warning Level Condition

$$p_i + s_i \geq p_{min} + \alpha s_{min} \qquad (2)$$

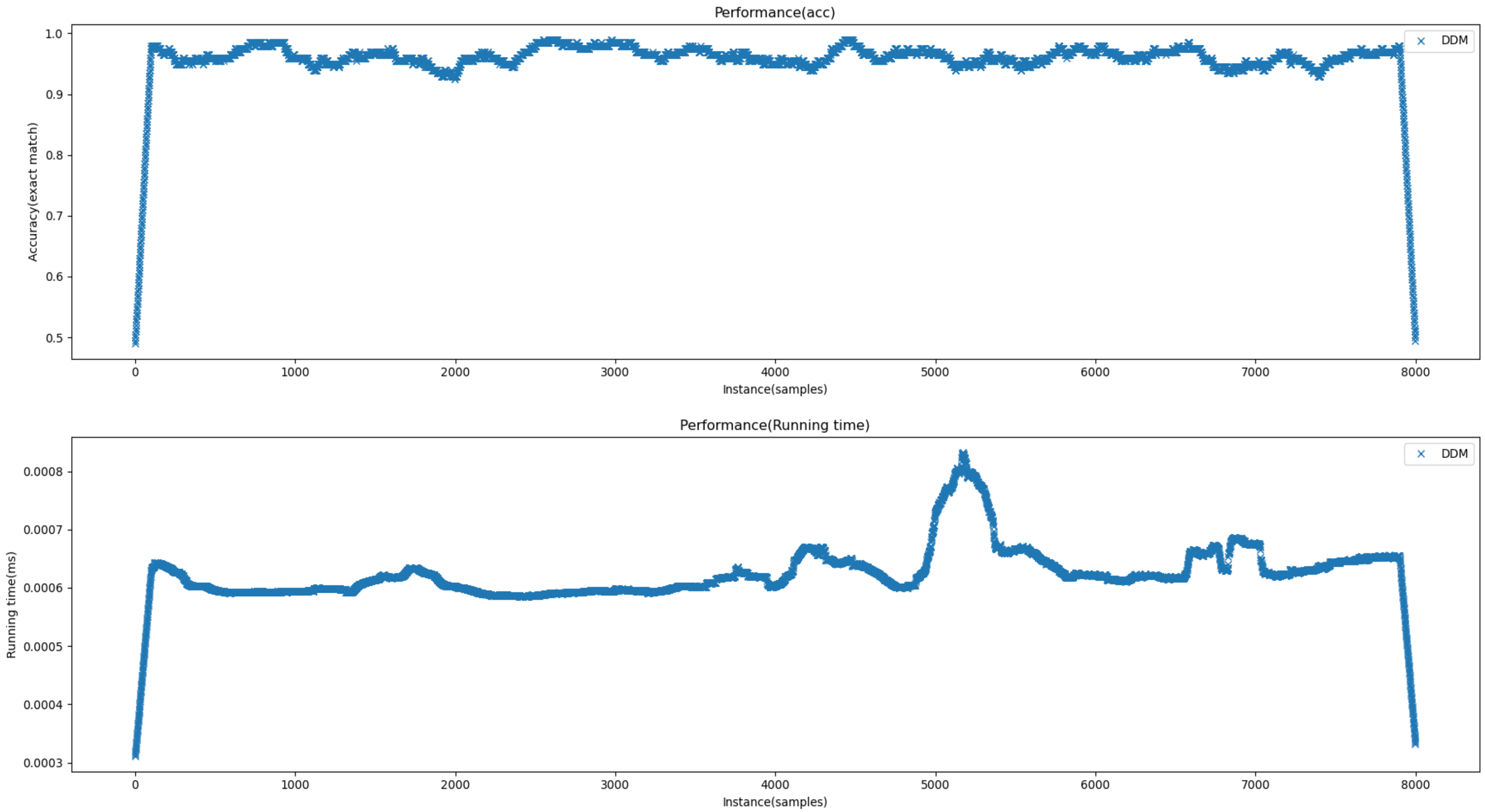
Alarm Level Condition

$$p_i + s_i \geq p_{min} + \beta s_{min} \qquad (3)$$

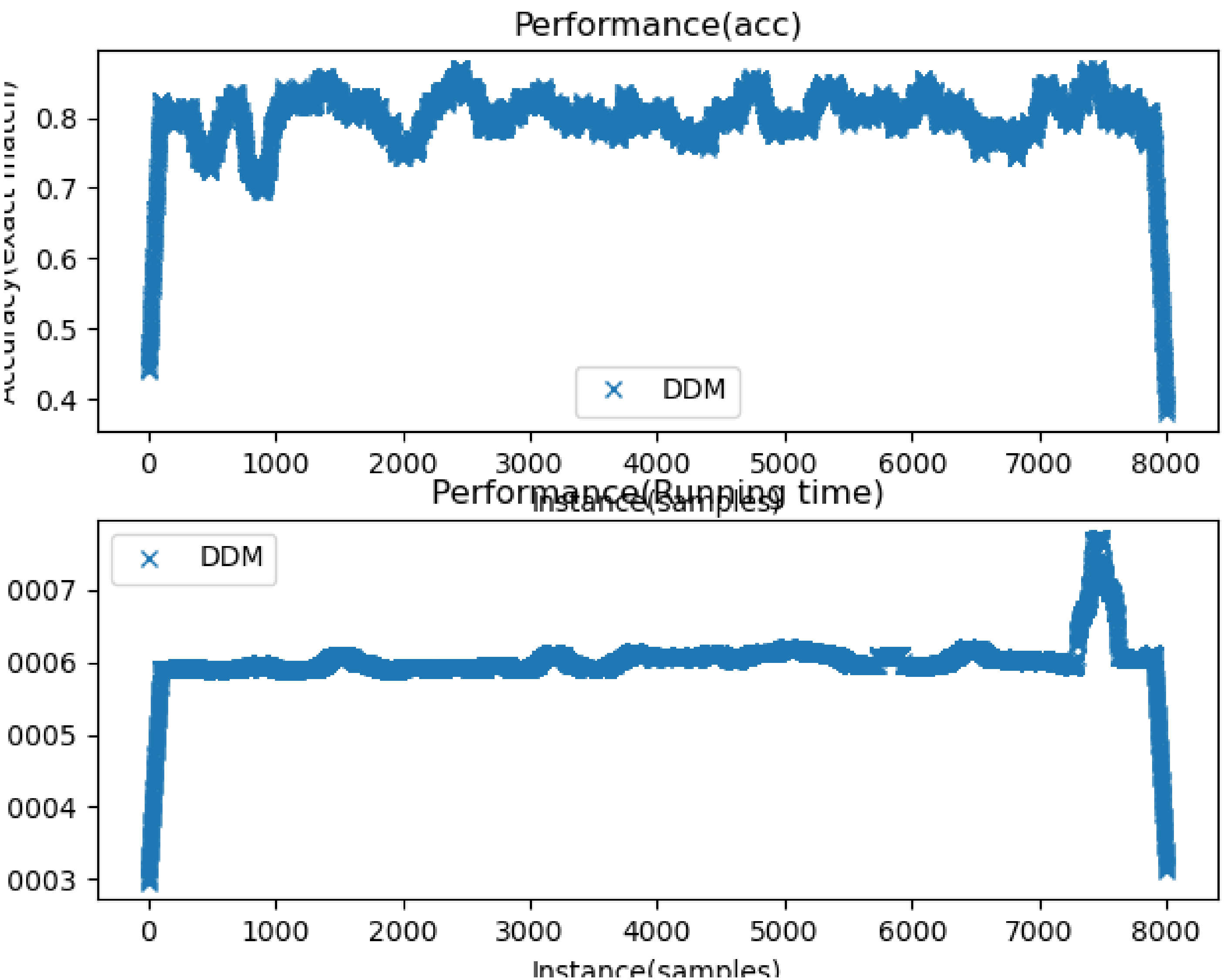
Dataset

- Four artificial datasets
- Balanced, contain 10,000 instances
- Datasets with noise, contain 10% examples with class noise.
- Allow us to evaluate the methods in terms of detection rate, detection delay and miss detection, besides prequential accuracy.
- The datasets are:Agrawal Gradual, Agrawal Abrupt, Sea Gradual, Sea Abrupt. The datasets are represented by nine features and divided into two classes.

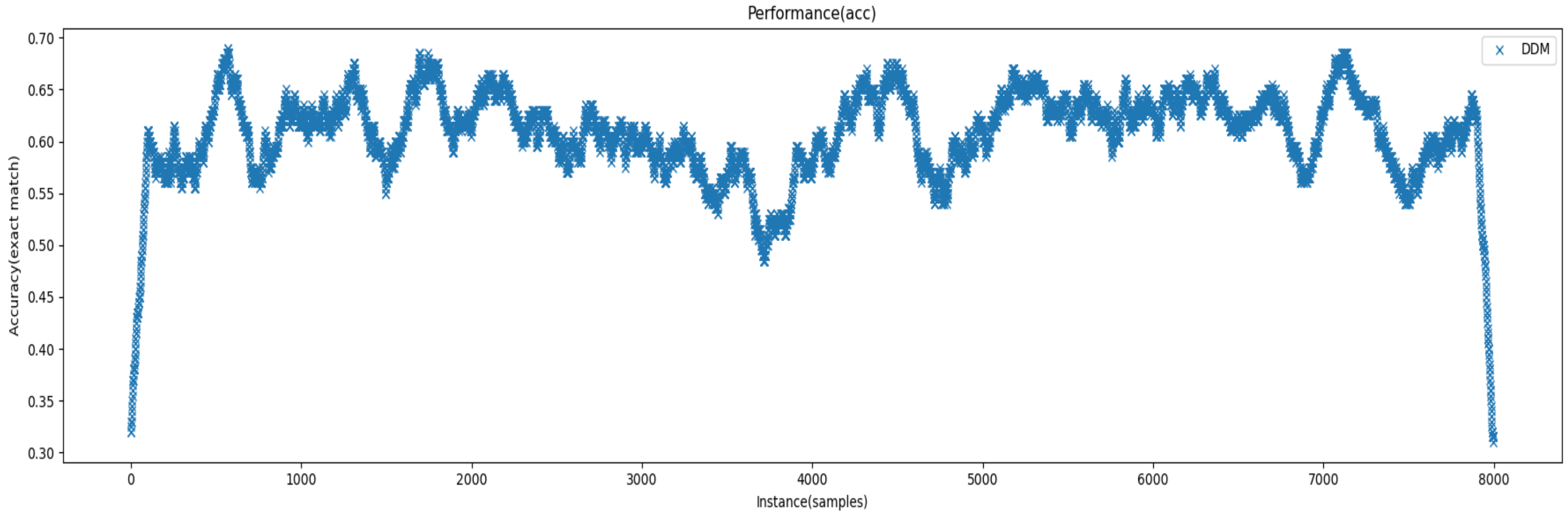
Agarwal Abrupt

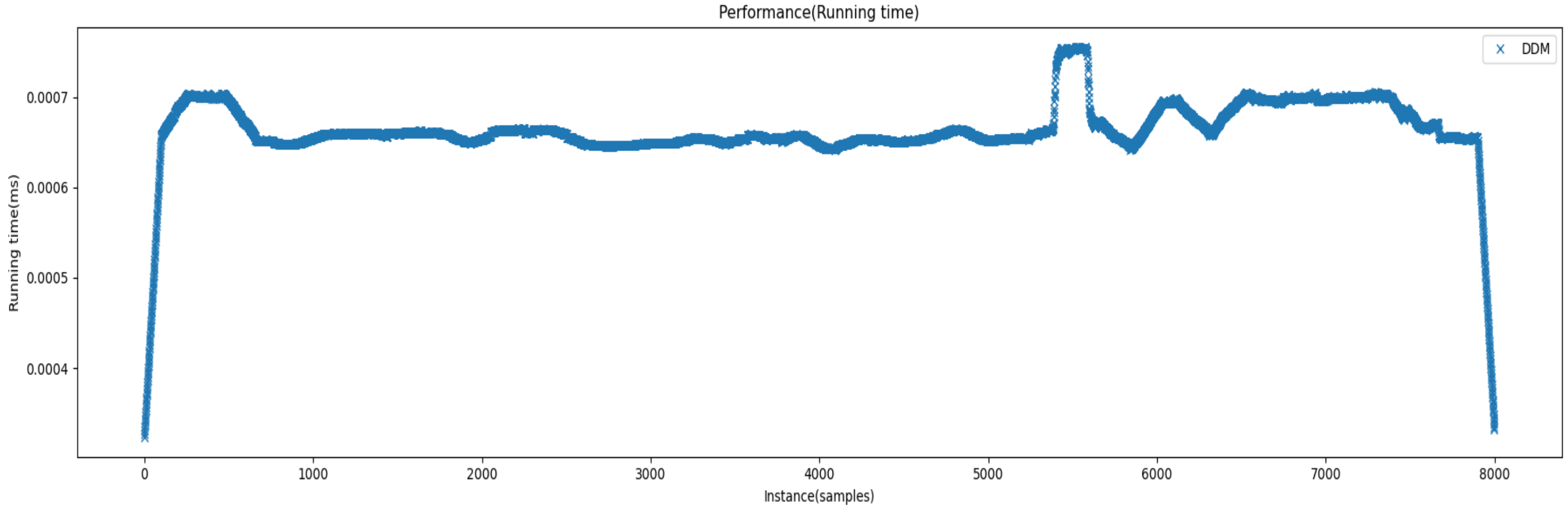


Agrawal Gradual

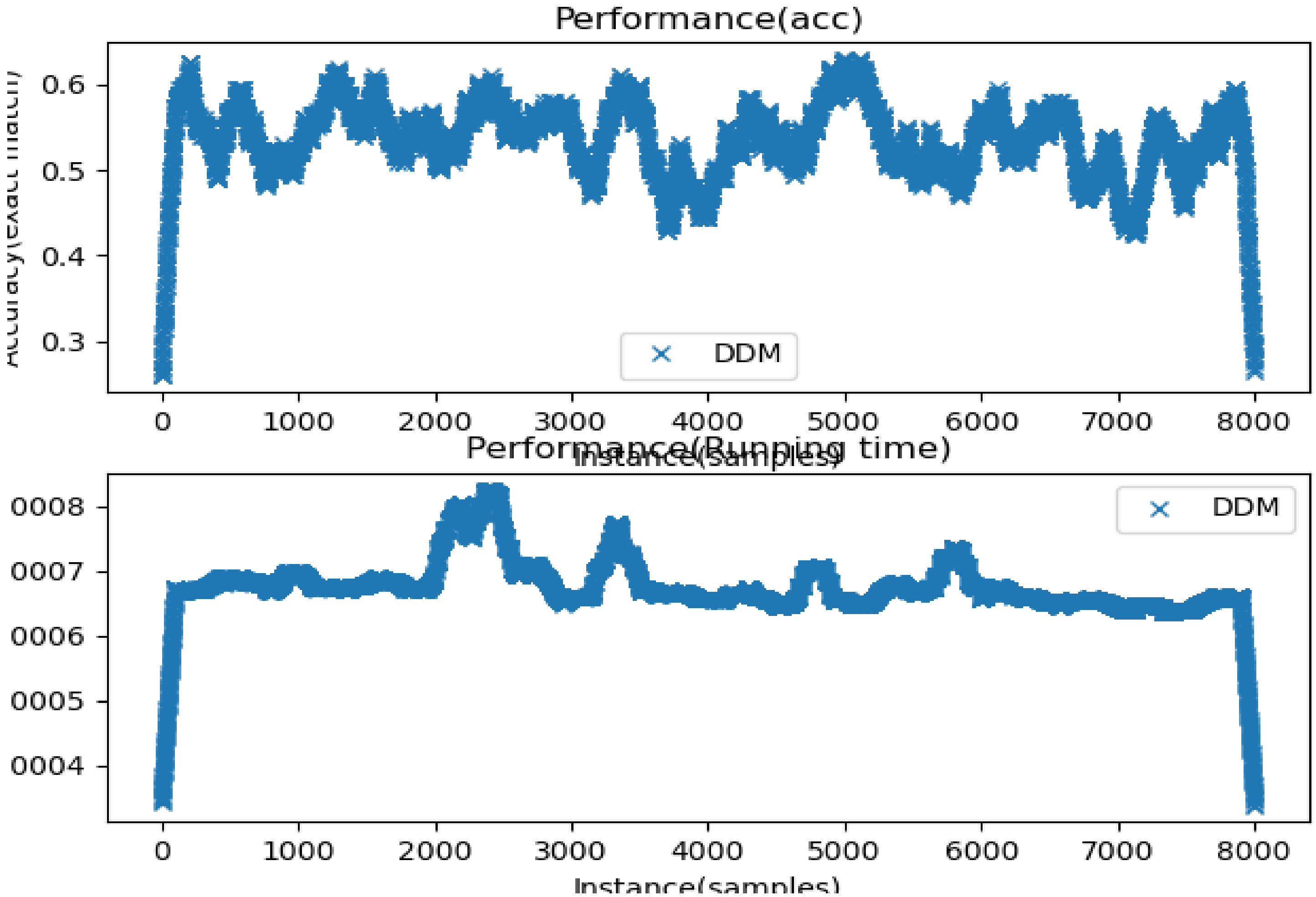


SEA Abrupt


Performance(acc)
DDM
Accuracy(exact match)
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0
1000
2000
3000
4000
5000
6000
7000
8000
Instance(samples)


Performance(Running time)
DDM
Running time(ms)
0.0007
0.0006
0.0005
0.0004
0
1000
2000
3000
4000
5000
6000
7000
8000
Instance(samples)

SEA Gradual


Performance(acc)
Accuracy(exact match)
0.6
0.5
0.4
0.3
DDM
0
1000
2000
3000
4000
5000
6000
7000
8000
Instance(samples)
Performance(Running time)
0008
0007
0006
0005
0004
DDM
0
1000
2000
3000
4000
5000
6000
7000
8000
Instance(samples)

Result Comparison

Dataset	Accuracy	Number of Detected Drifts
Agarwal Abrupt	96%	3
Agarwal Gradual	80%	11
Sea Abrupt	61%	17
Sea Gradual	54%	21

When the number of detected drift is high, the accuracy is lower.

Github page

Code	Add files via upload	now
dataset	Update and rename Agrawal to Readme.md	17 hours ago
README.md	Update README.md	18 hours ago

README.md

A Dicision-Based Dynamic Ensemble Selection Method for Concept Drift

cited paper: https://ieeexplore.ieee.org/document/8995320
R. A. S. Albuquerque, A. F. J. Costa, E. Miranda dos Santos, R. Sabourin and R. Giusti, "A Decision-Based Dynamic Ensemble Selection Method for Concept Drift," 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 2019, pp. 1132-1139, doi: 10.1109/ICTAI.2019.00158.

Abstract

The main task of this paper is concept drift. Concept drift will occur when data are continuously generated in streams, data and target concepts may change over time. For this problem `drift detector` is a commom solution, so the author proposed an online method which monitoring the stabilization of class distribution over time which named `Dynamic Ensemble Selection for Drift Detection(DESDD)`.

According to the idea of the author, the model should be able to estimate the class for each unknown instance, in order to raise the possibility of making a correct classification, the author then proposed an `ensemble-based method` which include **diverse population** of ensambles with different **member's diversity**, called `dynamic ensemble selection(DES)`, which elect a single ensemble that is probabily the best qualified to predict the class for given sample.

Proposed Method

DESDD is divided into four step:

Will add the link to the page in the presentation when all documents are completely uploaded

Thank you for your attention