GE 461 INTRODUCTION TO DATA SCIENCE
Analysis of Data Stream Classification Models on Different Datasets

Yagiz Yaman

Contents

In this project, multiple data stream classification models are applied on different datasets and the effect of concept drifts and the overall results are discussed.

OPTIONAL 1

Question

As an optional part, at the beginning of your report, you may have a related works section that covers data stream mining briefly with proper references.

Answer

With the advance in hardware and software systems, it becomes possible to process data which generated in high-speed due to an increase in the usage of technological devices and population. The term *data streams* is used to represent this rapidly generated data, for example, credit card transactions or phone calls in a city [1]. There are certain characteristics of *data streams* that differentiate it from static datasets:

- Potentially infinite number of observations,
- High rate of data arrival,
- Potential changes (concept drift) in data distribution during data stream process [3].

So, a data stream mining algorithms should process the vast incoming data and update its parameters to adapt to the changes as fast as possible [3]. Instance based classifiers, neural networks, Bayesian classifiers, and decision trees are standard machine learning methods for classifying data streams. Also, ensemble methods are quite promising to deal with a concept drift [3].

PART 4.1 - 4.2

The following classification models,

- Adaptive Random Forest (ARF)
- Streaming Agnostic Model with k-Nearest Neighbors (SAM-kNN)
- Streaming Random Patches (SRP)
- Dynamic Weighted Majority (DWM)

are applied on datasets,

- electric,
- spam,
- agrawal,
- sead.

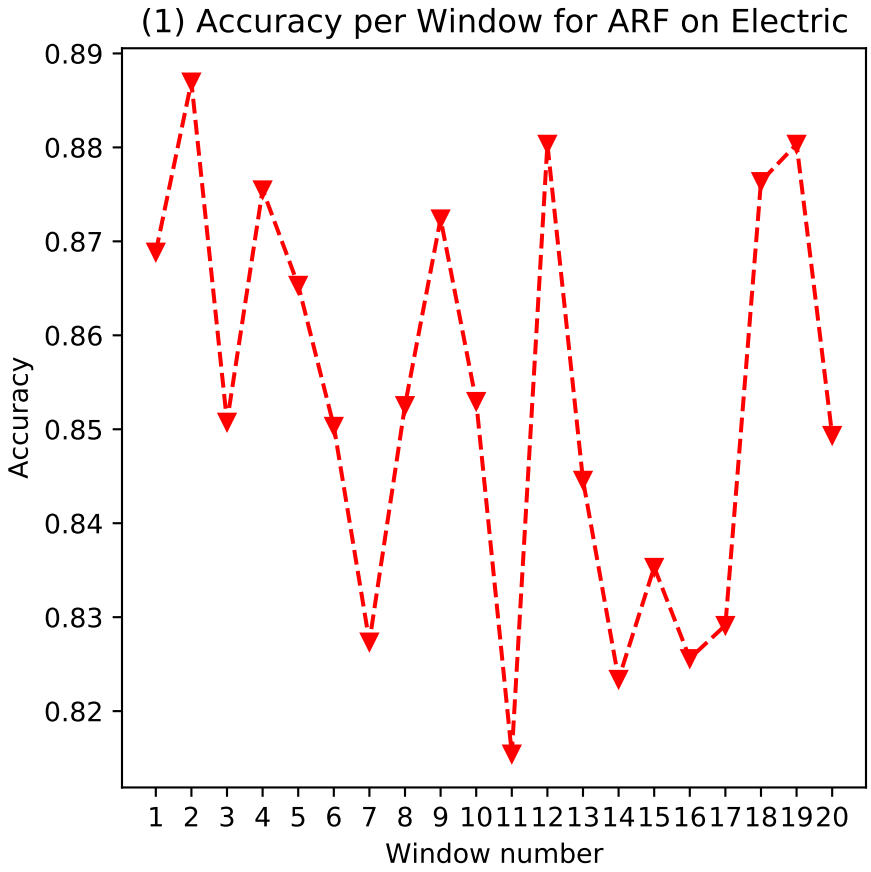
The run times and accuracies are provided.

DDM is used for a concept drift detection algorithm. The reason is it is good at capturing gradual changes. As the provided datasets do not have sudden changes, it works well.

The Interleaved Test-Then-Train approach is developed. As an evaluation metric, prediction accuracy is provided. Also, prequential accuracy plot is presented.

For electric dataset

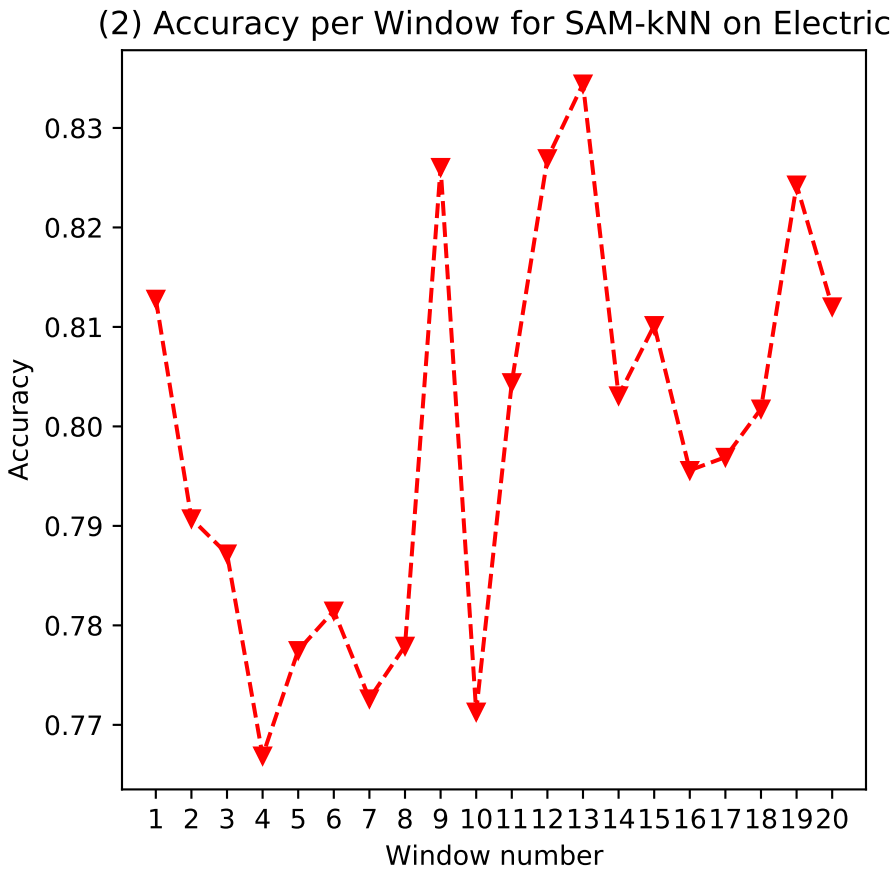
Adaptive Random Forest (ARF)




```
## Overall Accuracy = 85.32%
##  Run Time (sec) =  171
```

As **Table (1)** represents, the accuracy drops at specific points. The reason would be concept drifts. The ARF method achieves %85.65 overall accuracy. Its run time is 401 seconds which is relatively high compared to other methods.

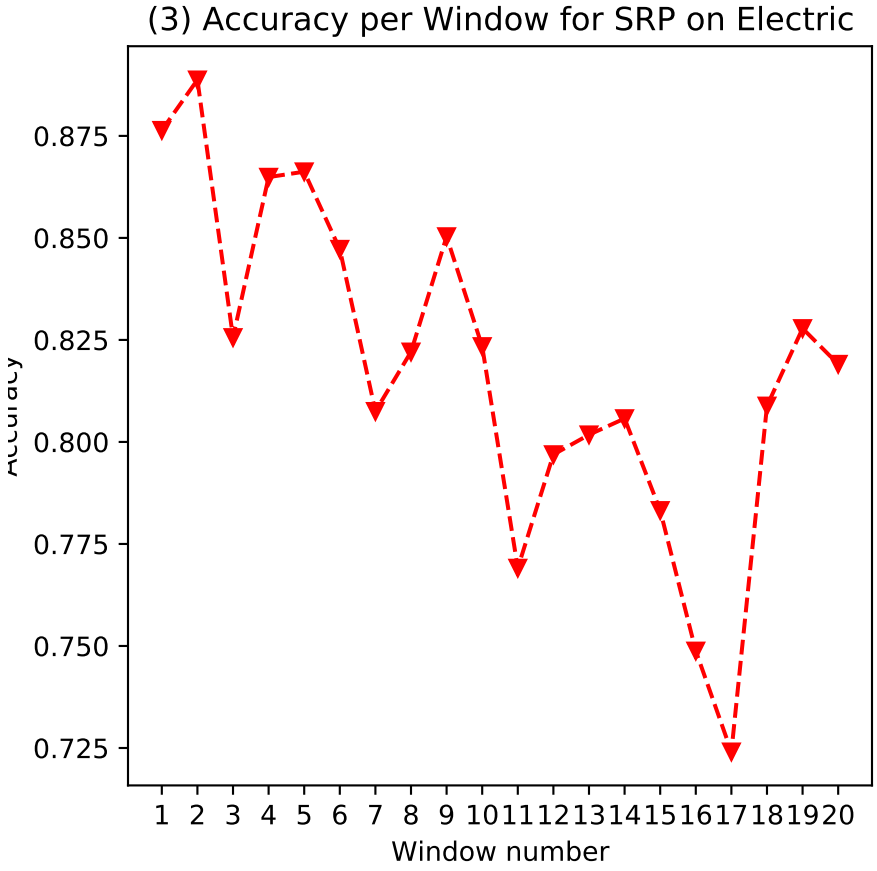
Streaming Agnostic Model with k-Nearest Neighbors (SAM-kNN)




```
## Overall Accuracy = 79.87%
##  Run Time (sec) =  49
```

Table (2) shows multiple concept drifts in the dataset. It is seen that ARF and SAM-kNN can catch different concept drifts as drops do not occur in the same intervals. The overall accuracy of SAMkNN is less than AFR, which is %85.65. Its run time is only 55 seconds, which is less than the AFR method.

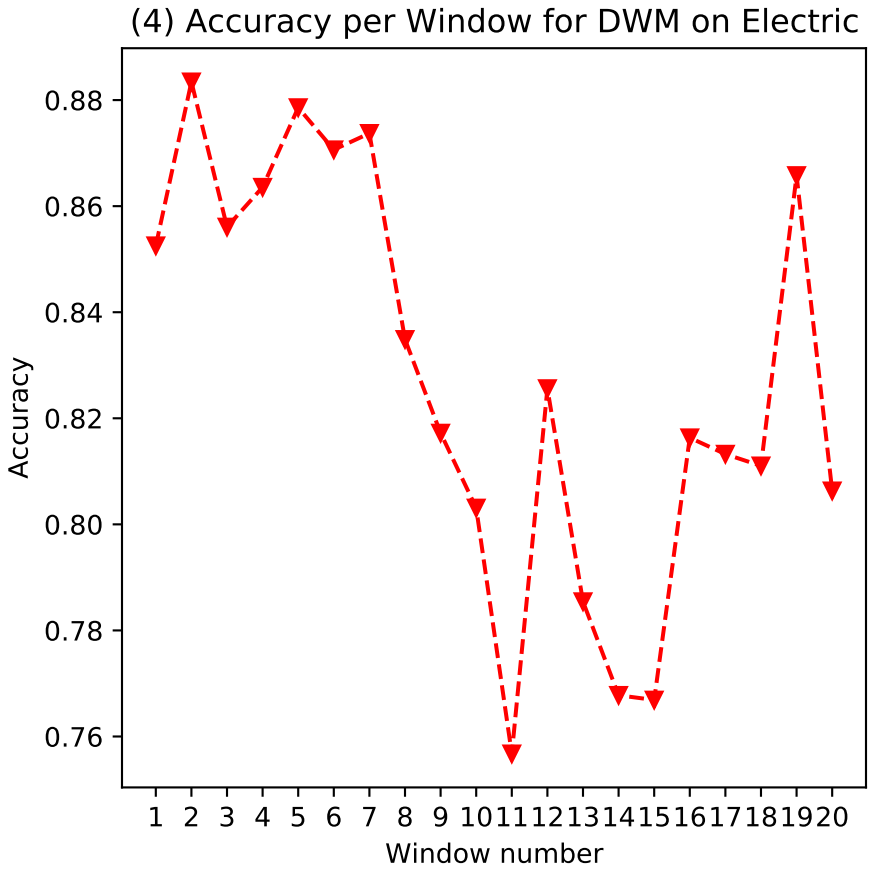
Streaming Random Patches (SRP)




```
## Overall Accuracy = 81.79%
##  Run Time (sec) =  240
```

There are similarities when we analyze the prequential accuracy plots of SRP and AFR. The accuracy drops around the same windows. So, these two methods are similar in responding to concept drifts. The overall accuracy is in the middle of AFR and SAMkNN, which is %81.79. Also, its run time is in between the AFR and SAMkNN, which is 259 seconds.

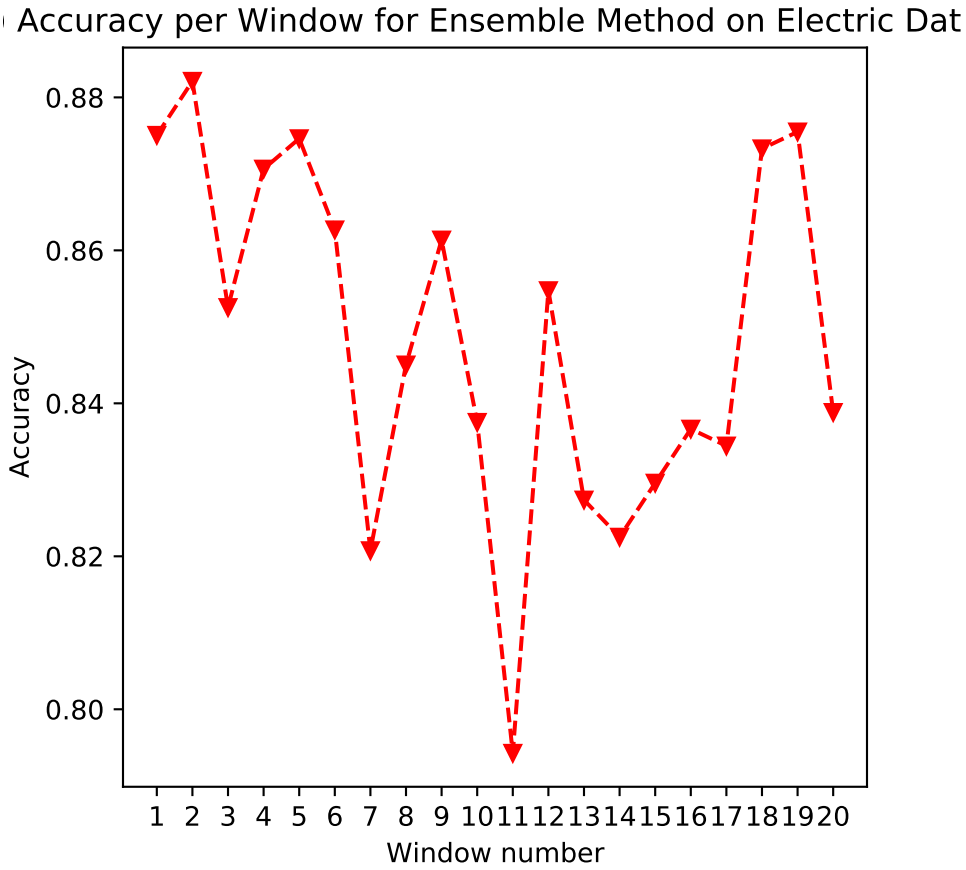
Dynamic Weighted Majority (DWM)




```
## Overall Accuracy = 82.74%
##  Run Time (sec) =  78
```

Table (4) looks similar to Table (2) in drops. There are steep reductions in accuracy from window to window. The overall accuracy is %82.74, and the run time is 63 seconds.

Ensemble Method

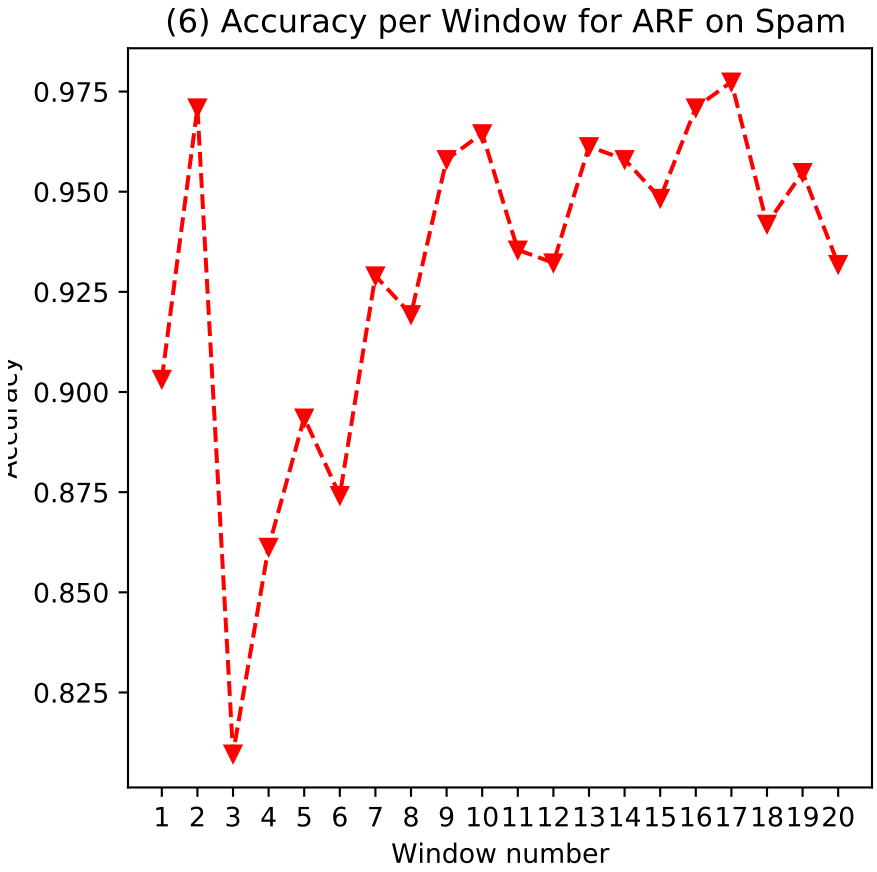



```
## Overall Accuracy = 84.85%
##  Run Time (sec) =  538
```

Table (5) represents the accuracy plot of the ensemble method. The overall accuracy of the ensemble method is %85, which is higher than the others, except AFR. Its run time is the highest as it uses the results of state-of-art approaches.

For spam dataset

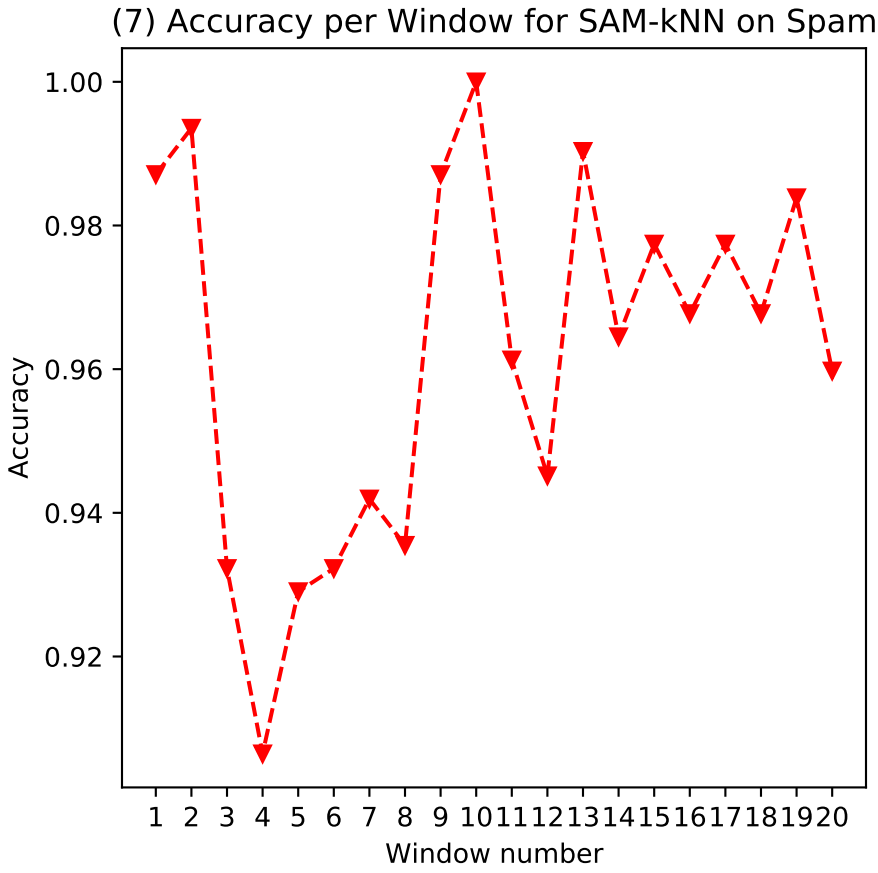
Adaptive Random Forest (ARF)




```
## Overall Accuracy = 92.98%
##  Run Time (sec) =  103
```

The spam dataset is the most extensive dataset regarding the number of features. Therefore it takes quite a lot of run time due to expensive computations. To prevent this, the number of estimators for AFR is set to 5. This would decrease the accuracy, but it is advantageous regarding run time. **Table (6)** shows concept drifts along with the dataset. The accuracy goes within the range of 0.8 and 0.98. The overall accuracy is %92.98, and the run time is 124 seconds.

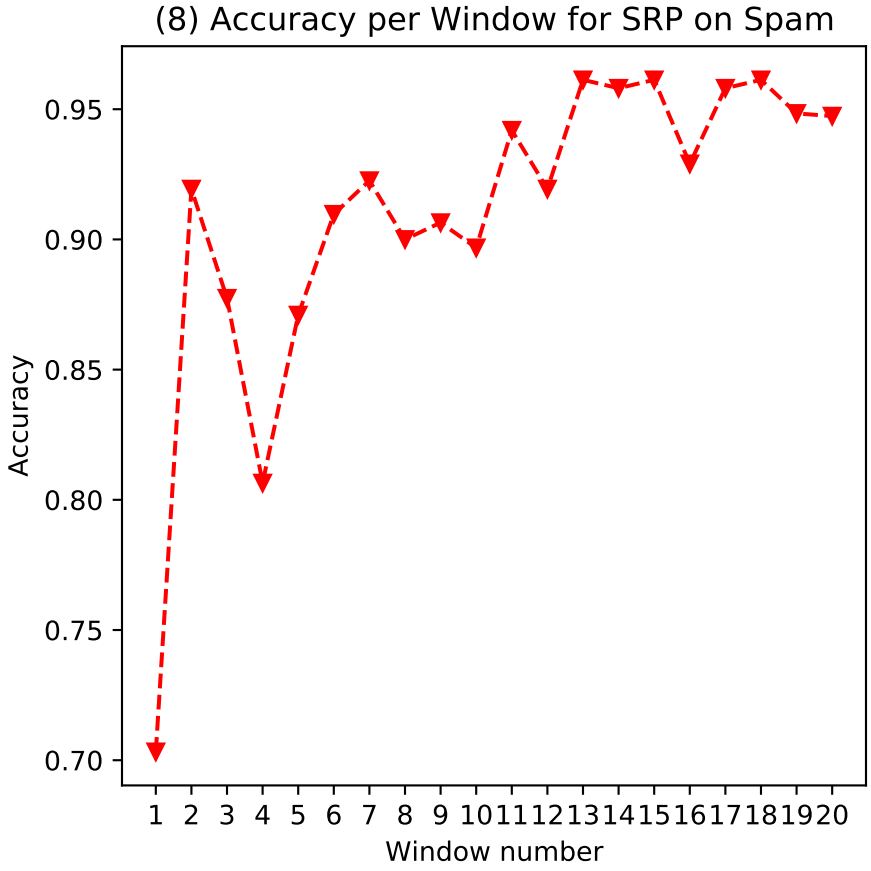
Streaming Agnostic Model with k-Nearest Neighbors (SAM-kNN)




```
## Overall Accuracy = 96.2%
##  Run Time (sec) =  68
```

SAMkNN classifier also works well in the spam dataset. The overall accuracy is %96.19. The run time is 81 seconds, which is quite promising. **Table (7)** represents the accuracy for each window. The accuracy of the ARF method drops significantly in window three; the same is true for SAMkNN. It would be a severe concept drift in the window three.

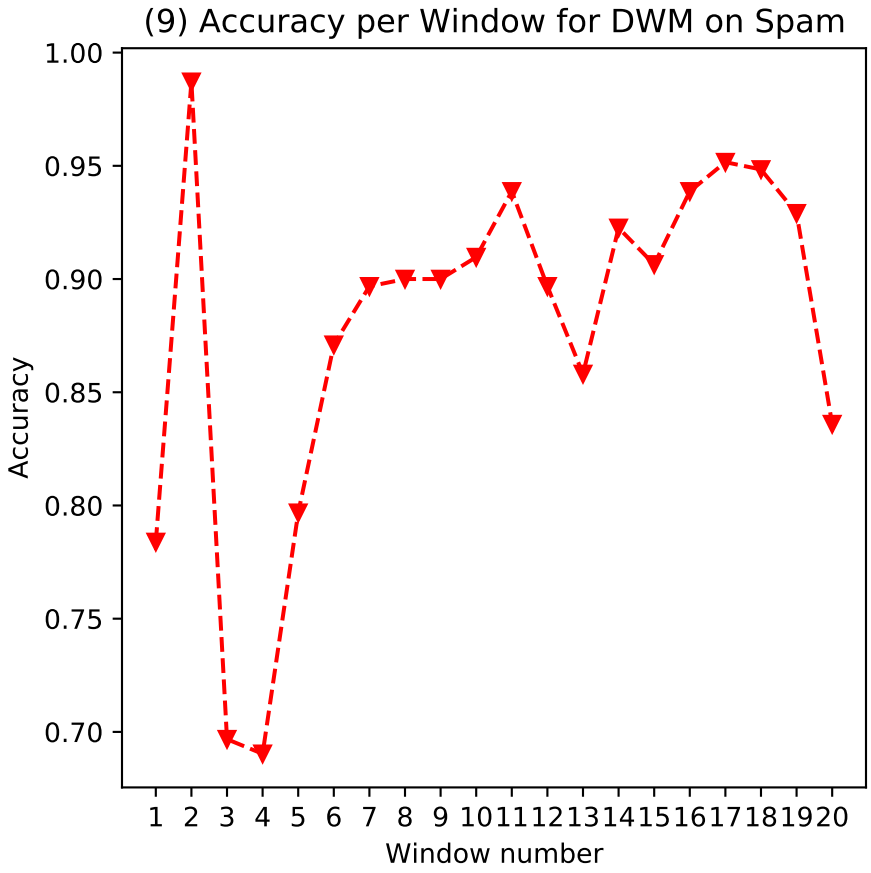
Streaming Random Patches (SRP)




```
## Overall Accuracy = 91.0%
##  Run Time (sec) =  589
```

Table (8) shows that SRP is quite good at responding to the concept drift in window three as it does not drop significantly as in the first two methods. The run time is relatively high, which is 688 seconds. It was even higher when the number of estimators was 10 (the default number). So, it is set to 3 to prevent extended run-time. The overall accuracy is %91.0.

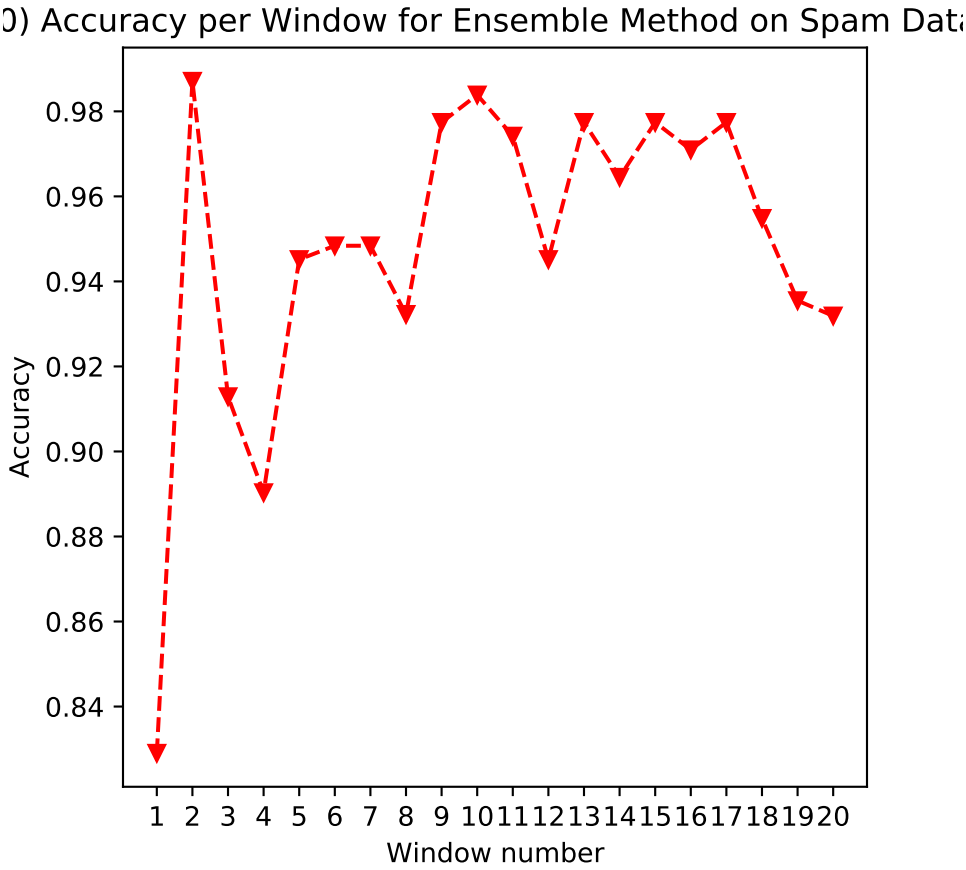
Dynamic Weighted Majority (DWM)




```
## Overall Accuracy = 87.78%
##  Run Time (sec) =  443
```

Table (9) shows the accuracy plot of the DWM method on the spam dataset. Its accuracy drops significantly at window 3. The DWM method is the worst one in accuracy compared to other state-of-art methods.

Ensemble Method

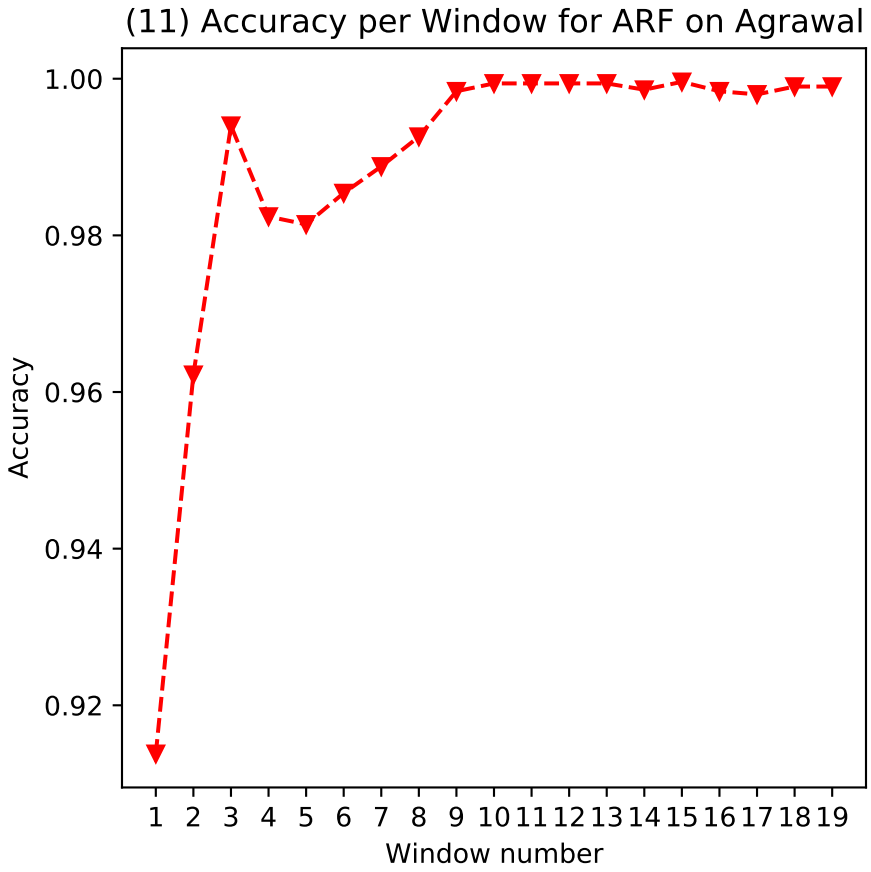



```
## Overall Accuracy = 94.82%
##  Run Time (sec) =  1203
```

Table (10) shows the accuracy plot of the ensemble method. The overall accuracy is %94.8.

For agrawal dataset

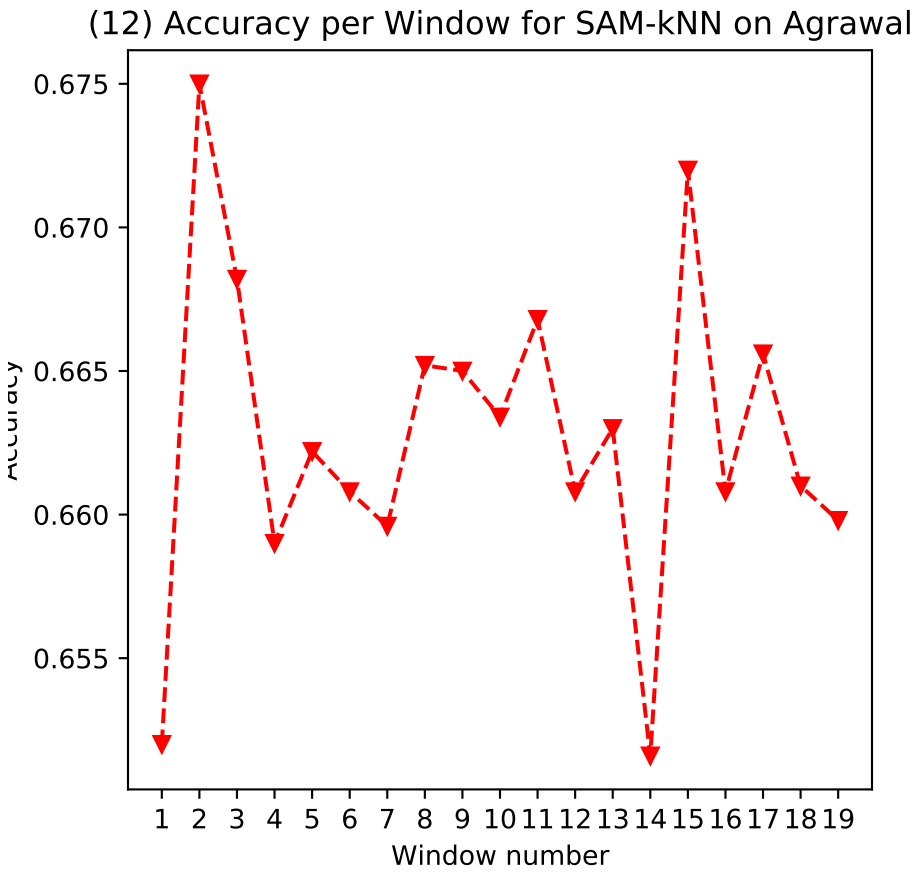
Adaptive Random Forest (ARF)




```
## Overall Accuracy = 98.94%
##  Run Time (sec) =  357
```

The ARF method is quite good at classifying the agrawal dataset, as its overall accuracy is %98.94. **Table (11)** indicates that the accuracy increases when the window number increases.

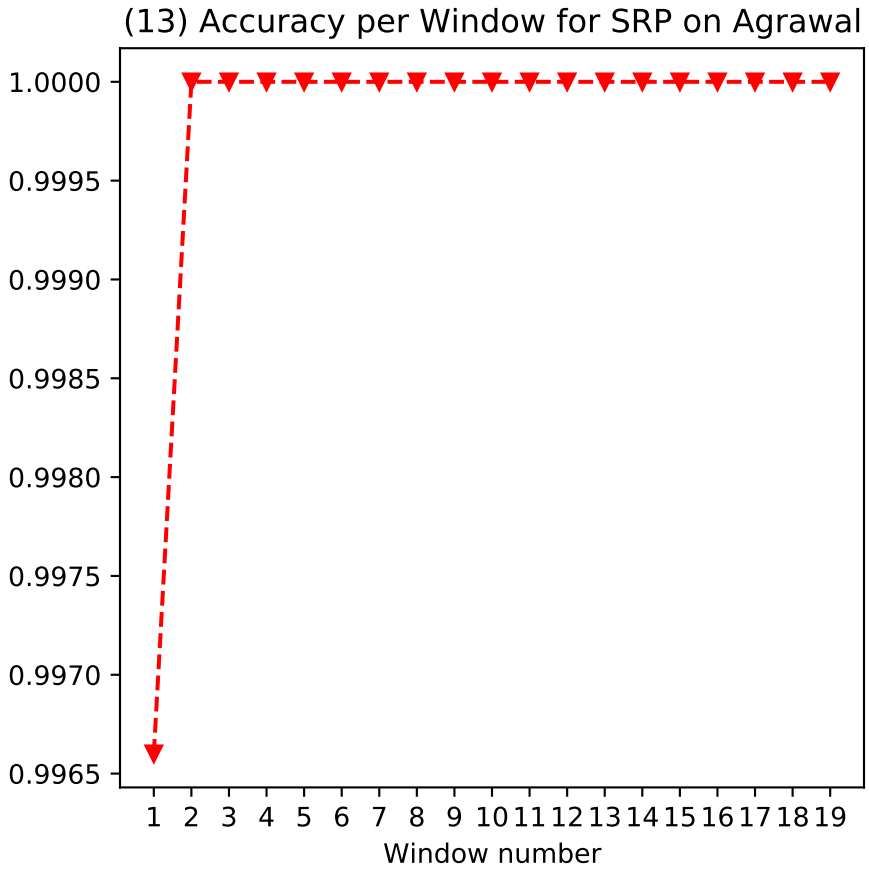
Streaming Agnostic Model with k-Nearest Neighbors (SAM-kNN)




```
## Overall Accuracy = 66.26%
##  Run Time (sec) =  80
```

The SAM-kNN works poorly on the agrawal dataset. Its accuracy is relatively low, which is %66.26. **Table (12)** represents the accuracy plot of the SAM-kNN method on the agrawal dataset for each window.

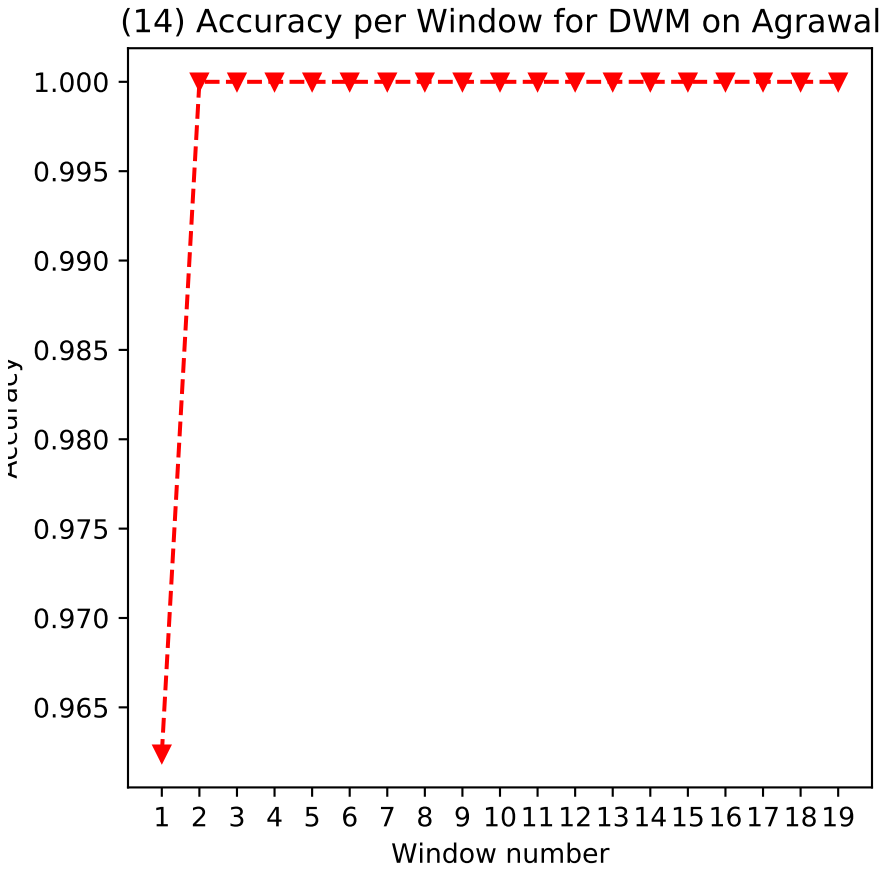
Streaming Random Patches (SRP)




```
## Overall Accuracy = 99.98%
##  Run Time (sec) =  557
```

Table (13) shows the accuracy plot of the SRP method on the agrawal dataset. The Overall accuracy is quite close to 100%. However, its run time is quite high.

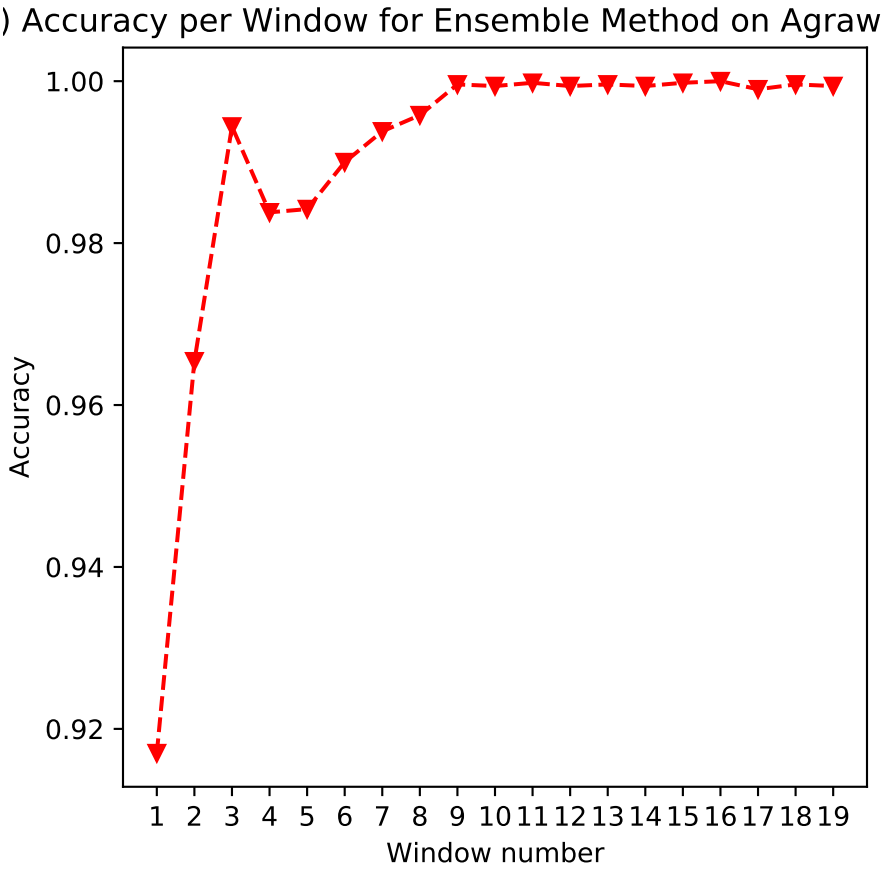
Dynamic Weighted Majority (DWM)




```
## Overall Accuracy = 99.81%
##  Run Time (sec) =  156
```

Table (14) represents that the DWM method predicts with nearly 100% accuracy, similar to SRP.

Ensemble Method

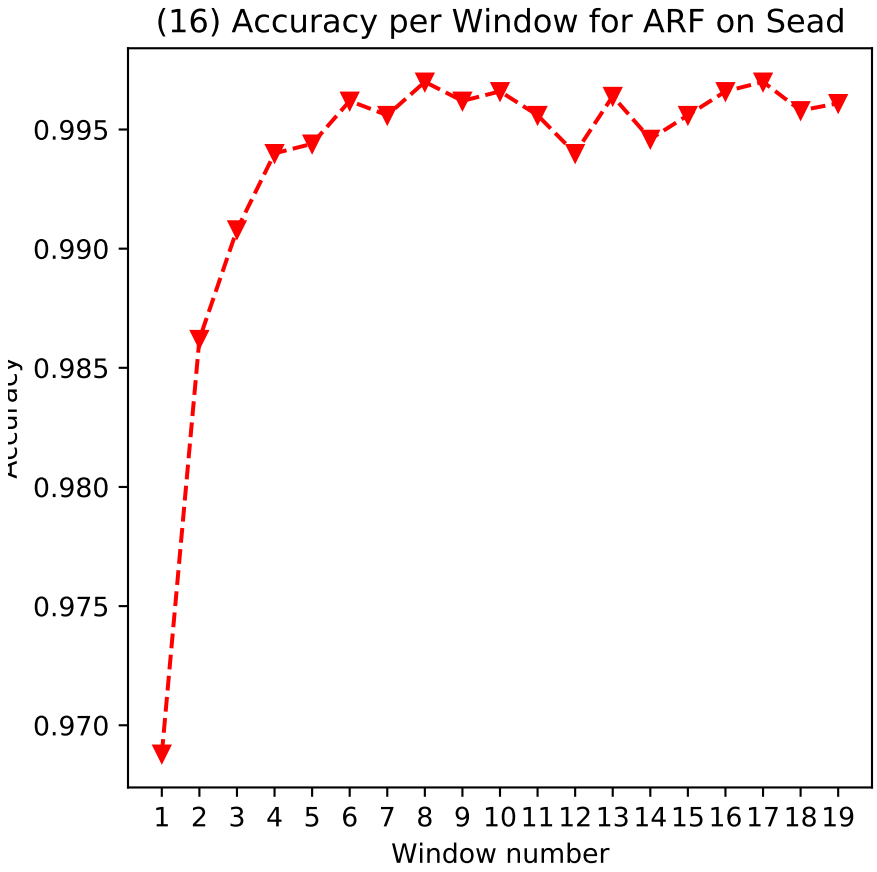



```
## Overall Accuracy = 99.09%
##  Run Time (sec) =  1150
```

As expected, **Table (15)** shows that the ensemble method is quite promising in classifying the agrawal dataset.

For sead dataset

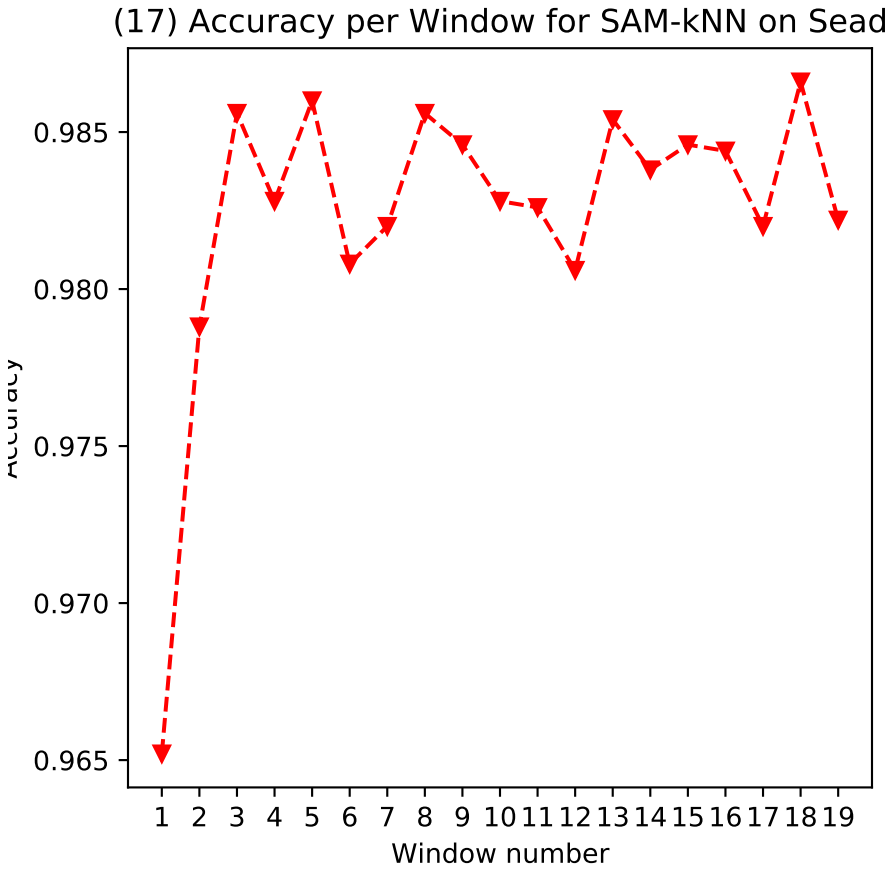
Adaptive Random Forest (ARF)




```
## Overall Accuracy = 99.37%
##  Run Time (sec) =  338
```

Table (16) represents that the ARF method is quite good at capturing the target in the sead dataset with 99.37% accuracy.

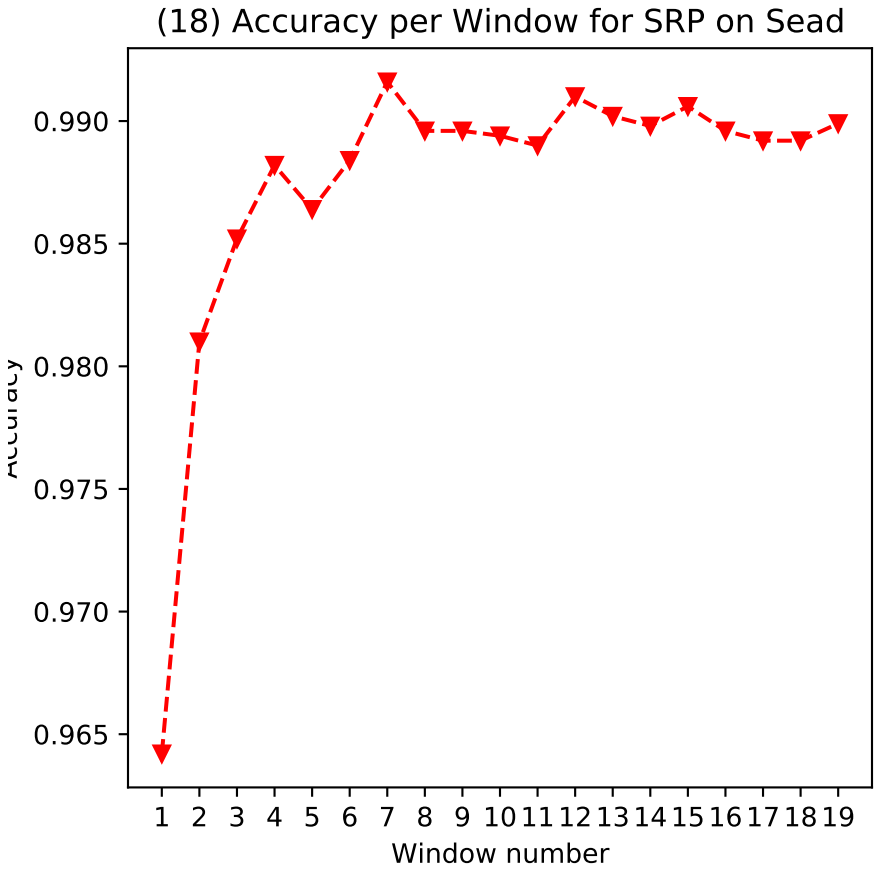
Streaming Agnostic Model with k-Nearest Neighbors (SAM-kNN)




```
## Overall Accuracy = 98.24%
##  Run Time (sec) =  359
```

The SAMkNN also performs well in the sead dataset, according to **Table (17)**. Its overall accuracy is 98.21%.

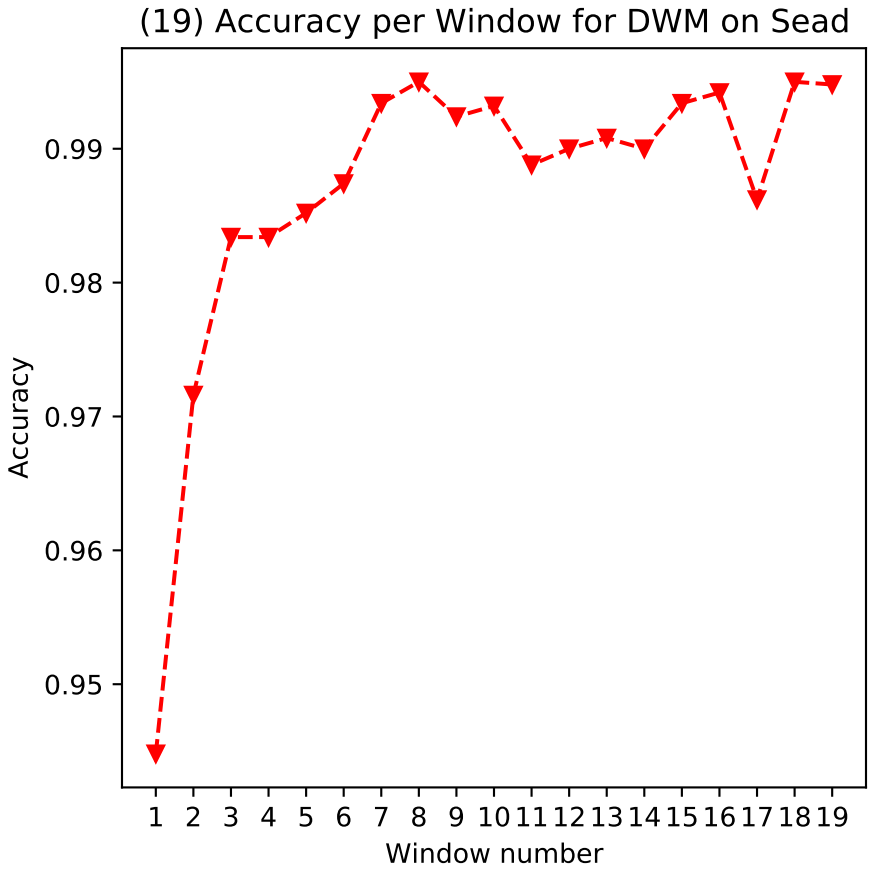
Streaming Random Patches (SRP)




```
## Overall Accuracy = 98.76%
##  Run Time (sec) =  223
```

According to **Table (18)**, the SRP method also gives promising results. Its Overall accuracy is %98.76.

Dynamic Weighted Majority (DWM)

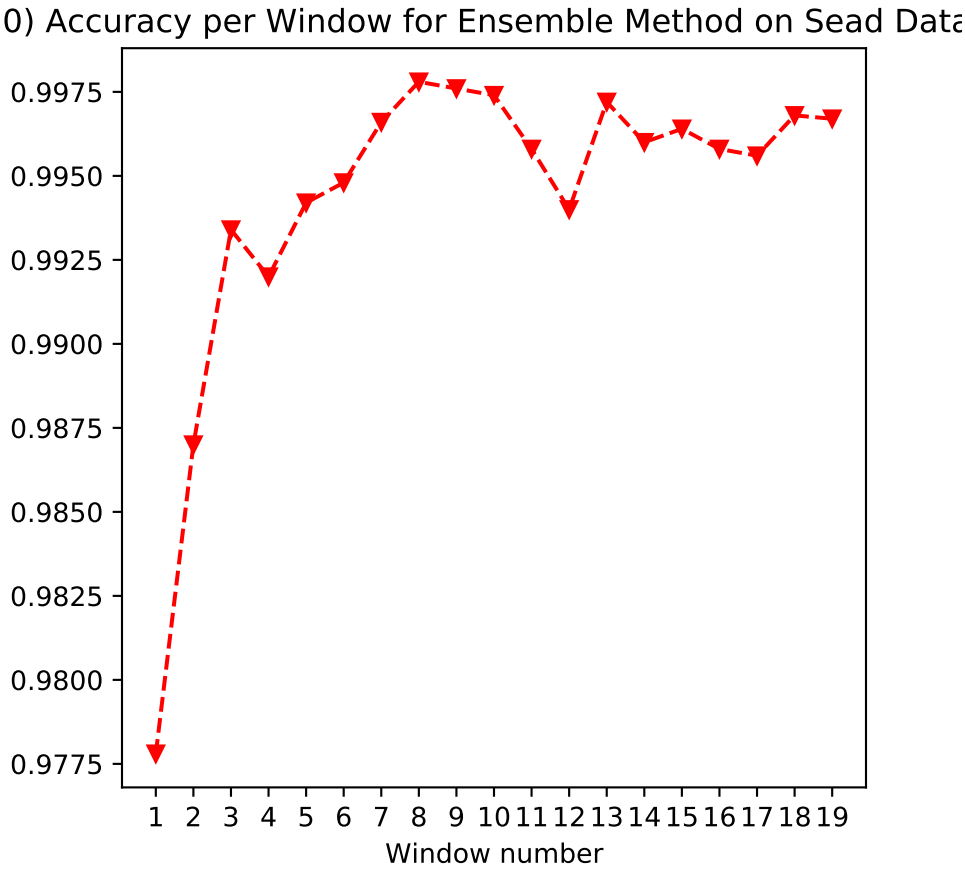



```
## Overall Accuracy = 98.74%
##  Run Time (sec) =  105
```

Like other methods, the DWM method performs well, as **Table (19)** represents. The overall accuracy is 98.74%.

Ensemble Method

```
## <string>:1: RuntimeWarning: More than 20 figures have been opened. Figures created through the pyplo
```




```
## Overall Accuracy = 99.45%
##  Run Time (sec) =  1025
```

Table (20) represents the accuracy plot. The overall accuracy of the ensemble method is quite promising, which is 99.41%.

5.1.

Question

How does your ensemble model perform compared to the state-of-the-art approaches in 4.1? What could be possible improvements for a more robust ensemble.

Answer

In general, the accuracy of the ensemble method is higher than the state-of-art methods. The robustness of the ensemble method can be increased by increasing the number of base models. Different models can capture various variability of the dataset, and as a result, this would increase the capability of the ensemble method. Also, different concept drift algorithms can be used in the base models as each model may perform better in specific concept drift.

5.2.

Question

Discuss your findings on the accuracy plots. What is inferred from the drops in the prequential accuracy plot?

Answer

When there is a concept drift, there is a drop in the prequential accuracy plot. It means the model cannot continue to perform with high accuracy. To mitigate this effect, a concept drift, DDM, is used. DDM is supposed to catch these concept drifts and make the model re-learn its parameters. So, because it updates the parameters, it is expected to increase accuracy after a while, as seen in the plots.

5.3.

Question

Please also include a paragraph that summarizes your findings in this assignment. What did you learn from this assignment?

Answer

In this assignment, I learned how to handle data streams where it is impossible to divide the data into train and test beforehand. I implemented five classification algorithms (ARF, SAM-kNN, SRP, DWM, and an ensemble model) with concept drift. The ensemble model, which takes the mode of the predictions from the first four models, generally outperforms the other models. We stated that the performance of the ensemble method could be improved by including different base models, using more estimators, and applying different concept drift mechanisms to base models. I analyzed the prequential accuracy plots and stated that there are indications of possible concept drifts. I concluded that concept drift decreases the performance of a model for a while.

Overall, the assignment taught me the concept of the data stream and the effect of incorporating concept drift into classification algorithms.

OPTIONAL 2

Question

Another optional part is comparison of the effectiveness of the methods using statistical tests. The design and administration of these tests should be decided by you by looking at the available papers in literature.

Answer

In some cases, accuracy might be misleading due to the dataset. For example, if a dataset mainly consists of one class, the model would be biased, and it might perform poorly on a balanced dataset. For example, if a dataset has only class "1", a model which always gives the output "1" will result in 100% accuracy, but this model is not a proper model. So, if the classifier is trained on an unbalanced dataset, other statistical tests, such as precision, recall, f-measure, and specificity, should be applied [4].

So, the corresponding metrics of each model on the all datasets are provided below.

```
## Confusion Matrix for the:ARF
## [[126082   3544]
##  [  5237 116662]]
## Precision for the ARF = 0.97
## Recall for the ARF = 0.96
```

```
## F-measure for the ARF = 0.96
## Specificity for the ARF = 0.97

## Confusion Matrix for the:SAMkNN
## [[122339   7287]
##  [ 37569  84330]]
## Precision for the SAMkNN = 0.92
## Recall for the SAMkNN = 0.69
## F-measure for the SAMkNN = 0.79
## Specificity for the SAMkNN = 0.94

## Confusion Matrix for the:SRP
## [[124695   4931]
##  [  5137 116762]]
## Precision for the SRP = 0.96
## Recall for the SRP = 0.96
## F-measure for the SRP = 0.96
## Specificity for the SRP = 0.96

## Confusion Matrix for the:DWM
## [[125030   4596]
##  [  5431 116468]]
## Precision for the DWM = 0.96
## Recall for the DWM = 0.96
## F-measure for the DWM = 0.96
## Specificity for the DWM = 0.96

## Confusion Matrix for the:Ensemble
## [[127377   2249]
##  [  6397 115502]]
## Precision for the Ensemble = 0.98
## Recall for the Ensemble = 0.95
## F-measure for the Ensemble = 0.96
## Specificity for the Ensemble = 0.98
```

The **ensemble model** achieved the highest precision (98%), indicating that it correctly classified a high proportion of 1's out of all the targets it classified as 1's. However, its recall (94.7%) was slightly lower than the ARF, SRP and DWM, indicating that it missed a few 1's. Also, the worst model is **SAMkNN** in all metrics.

REFERENCES

[1] M. M. Gaber, "Advances in Data Stream Mining," *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 79-85, 2011. doi:10.1002/widm.52

[2] L. Rutkowski, M. Jaworski, and P. Duda, *Stream Data Mining: Algorithms and Their Probabilistic Properties*, Cham: Springer International Publishing, 2020.

[3] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, Springer, 2005.