



Employing chunk size adaptation to overcome concept drift


Jędrzej Kozal

(Wrocław University of Science and Technology, Wrocław, Poland
 <https://orcid.org/0000-0001-7336-2561>, jedrzej.kozal@pwr.edu.pl)

Filip Guzy

(Wrocław University of Science and Technology, Wrocław, Poland
 <https://orcid.org/0000-0002-8619-7080>, filip.guzy@pwr.edu.pl)

Michał Woźniak

(Wrocław University of Science and Technology, Wrocław, Poland
 <https://orcid.org/0000-0003-0146-4205>, michal.wozniak@pwr.edu.pl)

Abstract: Modern analytical systems must process streaming data and correctly respond to data distribution changes. The phenomenon of changes in data distributions is called *concept drift*, and it may harm the quality of the used models. Additionally, the possibility of *concept drift* appearance causes that the used algorithms must be ready for the continuous adaptation of the model to the changing data distributions. This work focuses on non-stationary data stream classification, where a classifier ensemble is used. To keep the ensemble model up to date, the new base classifiers are trained on the incoming data blocks and added to the ensemble while, at the same time, outdated models are removed from the ensemble. One of the problems with this type of model is the fast reaction to changes in data distributions. We propose the new *Chunk Adaptive Restoration* framework that can be adapted to any block-based data stream classification algorithm. The proposed algorithm adjusts the data chunk size in the case of *concept drift* detection to minimize the impact of the change on the predictive performance of the used model. The experimental research, backed up with the statistical tests, has proven that *Chunk Adaptive Restoration* significantly reduces the model's restoration time.

Keywords: Data stream, Data stream mining, Continual learning, Pattern classification, Concept drift, Block-based data processing

Categories: H.2.8, I.2.6, I.5.0, I.5.2

DOI: 10.3897/jucs.80735

1 Introduction

Data stream mining focuses on the knowledge extraction from streaming data, mainly for the predictive model construction aimed at assigning arriving instances to one of the predefined categories. This process is characterized by additional difficulties caused by the data distribution evolution. It is visible in many practical tasks as spam detection, where the spammers still change the message format to cheat anti-spam systems. Another example is medical diagnostics, where new SARS-CoV-2 mutations may cause different symptoms, which forces doctors to adapt and improve diagnostic methods [Harvey et al., 2021].

The phenomenon mentioned above is called *concept drift*, and its nature can vary due to both the character and the rapidity. It forces classification models to adapt to new data characteristics and forget previously known, useless concepts. An important characteristic of such systems is their reaction to the *concept drift* phenomenon, i.e., how much predictive performance deteriorates when it occurs and when the classification system will obtain the approved predictive quality for the new concept. We should also consider another limitation: the classification system should be ready to classify incoming objects immediately, and dedicated computing and memory resources are limited.

Data processing models used by stream data classification systems can be roughly divided into two categories: online (object by object) processing (online learners) or block-based (chunk by chunk) data processing (block-based learners) [Krawczyk et al., 2017]. Online learners require model parameters to be updated when a new object appears, while the block-based method requires updates once per batch. The advantage of online learners is their fast adaptation to *concept drift*. However, in many practical applications, the effort of necessary computation (related to updating models after processing each object) is unacceptable. The model update can require many operations that involve changing data statistics, updating the model's internal structure, or learning a new model from scratch. These requirements can become prohibitive for high-velocity streams. Hence, more popular is block-based data processing, which requires less computational effort. However, it limits the model's potential for quick adaptation to changes in data distribution and fast restoration of performance after *concept drift*. Consequently, a significant problem is the proper selection of the chunk size. Smaller data block size results in faster adaptation. However, it increases the overall computing load. On the other hand, larger data chunks require less computation but result in a lower adaptive capacity of the classification model. Another valid consideration is the impact of chunk size on prediction stability. Models trained on smaller chunks typically have more significant prediction variance, while models trained with larger chunks tend to have more stable predictions when the data stream is stationary. If *concept drift* occurs, a larger chunk size increases the probability that the data from different concepts will be placed in the same batch. Hence, selecting the chunk size is a trade-off encompassing computation power, adaptation speed, and predictions variance.

The trade-off described above includes features that are equally desired in many applications. Especially both the consumption of computational power and the speed of adaptation are important when processing large data streams. We propose a new method that alleviates the downfalls of choosing between small or large chunk sizes by dynamically changing the current batch size. More precisely, our work introduces the *Chunk-Adaptive Restoration* (CAR), a framework based on combined drift and stabilization detection techniques that adjusts the chunk sizes during the *concept drift*. This approach slightly redefines the previous problem based on the observation that for many practical classification tasks, a period of changes in data distributions is followed by stabilization. Hence, we propose that when the *concept drift* occurs, the model should be quickly upgraded, i.e., the data should be processed in small chunks, and during the stabilization period, the data block size may be extended. The advantage of the proposed method is its universality and the possibility of using it with various chunk-based data stream classifiers.

This work offers the following contributions:

- Proposing the *Chunk-Adaptive Restoration* framework to empower fluent restoration after *concept drift* appearance.
- Formulating the *Variance-based Stabilization Detection Method*, a technique com-

plementary to all *concept drift* detectors that simplifies chunk size adaptation and metrics calculation.

- Employing *Chunk-Adaptive Restoration* for the adaptive data chunk size setting for selected state-of-the-art algorithms.
- Introducing a new stream evaluation metric, *Sample Restoration*, to show the gains of the proposed methods.
- Evaluating the proposed approach based on various synthetic and real data streams and a detailed evaluation of its usefulness for the selected state-of-the-art methods.

2 Related works

This section provides a review of the related works. Firstly, we will discuss challenges specific to the learning from non-stationary data streams. Next, we discuss different methods of processing data streams. Following, we describe existing drift detection algorithms and ensemble methods. We continue by reviewing existing evaluation protocols and computational and memory requirements. We conclude this section by providing examples of other data stream learning methods that employ variable chunk size.

2.1 Challenges related to data stream mining

A data stream is a sequence of objects described by their attributes. In the case of a classification task, each learning object should be labeled. The number of items may be vast, potentially infinite. Observations in the stream may arrive at different times, and the time intervals between their arrival could vary considerably. The main differences between analyzing data streams and static datasets include [Bifet et al., 2018]:

- No one can control the order of incoming objects.
- The computation resources are limited, but the analyzer should be ready to process the incoming item in a reasonable time.
- The memory resources are also limited, but the data stream size may be huge or even infinite, which causes memorizing all the items impossible.
- Data streams are susceptible to change, i.e., data distributions may change over time.
- The labels of arriving items are not free, for some cases impossible to get, or available with delay (e.g., in banking for credit approval task after a few years).

The canonical classifiers usually do not consider that the probabilistic characteristics of the classification task may evolve [Duda et al., 2001]. Such a phenomenon is known as *concept drift* [Widmer and Kubat, 1996], and a few *concept drift* taxonomies have been proposed. The most popular consider how rapid the drift is, then we can distinguish sudden drift and incremental one. An additional difficulty is a case when, during the transition between two concepts, objects from two different concepts appear for some time simultaneously (gradual drift). We can also take into consideration the influence of the probabilistic characteristics on the classification task [Joao et al., 2014]:

- Virtual *concept drift* does not impact the decision boundaries but affects the probability density functions [Widmer and Kubat, 1993], and Widmer and Kubat [Widmer and Kubat, 1996] imputed it rather to incomplete data representation than to the true changes in concepts,
- Real *concept drift* affects the *posterior* probabilities and may impact the unconditional probability density function [Widmer and Kubat, 1996].

2.2 Methods for processing data streams

The data stream can be divided into small portions of the data called data chunks. This method is known as batch-based or chunk-based learning. Choosing the proper size of the chunk is crucial because it may significantly affect the classification [Junsawang et al., 2019]. Unfortunately, the unpredictable appearance of the *concept drift* makes it difficult. Several approaches may help overcome this problem, e.g., using different windows for processing data [Lazarescu et al., 2004] or adjusting chunk size dynamically [Widmer and Kubat, 1996]. Unfortunately, most chunk-based classification methods assume that the size of the data chunk is priorly set and remains unchanged during the data processing.

Instead of chunk-based learning, the algorithm can learn incrementally (online) as well. Training examples arrive one by one at a given time and are not kept in memory. The advantage of this solution is the need for small memory resources. However, the effort of necessary computation related to updating models after processing each individual object is unacceptable, especially in the high-velocity data streams, i.e., Internet of Things (IoT) applications.

When processing a non-stationary data stream, we can rely on a drift detector to point moments when data distribution has changed and take appropriate actions. The alternative is to use inherent adaptation properties of models (update & forget). In the following subsection, we will discuss both of these approaches.

2.3 Drift detection methods

A drift detector is an algorithm that can inform any changes within data stream distributions. The data labels or a classifier's performance (measured using any metric, such as accuracy) is required to detect a real *concept drift* [Sobolewski and Wozniak, 2013]. We have to realize that drift detection is a non-trivial task. The detection should be done as quickly as possible to replace an outdated model and minimize restoration time. On the other hand, false alarms are unacceptable, as they will lead to an incorrect model adaptation and resource spending where there is no need for it [Gustafsson, 2000]. *Drift Detection Method* (DDM) [Gama et al., 2004] is one of the most popular detectors that incrementally estimates an error of a classifier. Because we assume the classifier training method's convergence, the error should decrease with the appearance of subsequent learning objects [Raudys, 2014]. If the reverse behavior is observed, we may suspect a change of probability distributions. DDM uses the *three-sigma rule* to detect a drift. *Early Drift Detection Method* (EDDM) [Baena-Garcia et al., 2006] is an extension of DDM, where the window size selection procedure is based on the same heuristics. Additionally, the *distance error rate* is used instead of the classifier's error rate. Blanco et al. [Blanco et al., 2015] proposed very interesting drift detectors that use the non-parametric estimation of classifier error employing Hoeffding's and McDiarmid's inequalities.

2.4 Ensemble methods

One of the most promising data stream classification research directions, which usually employs chunk-based data processing, is the classifier ensemble approach [Krawczyk et al., 2017]. Its advantage is that the classifier ensemble can easily adapt to the *concept drift* using different updating strategies [Kuncheva, 2004]:

- *Dynamic combiners* – individual classifiers are trained in advance, and they are not updated anymore. The ensemble classifier adapts to changing data distribution by changing the combination rule parameters.
- *Updating training data* – incoming examples are used to retrain component classifiers (e.g., online bagging [Oza and Tumer, 2008]).
- *Updating ensemble members* [Bifet et al., 2009, Rodríguez and Kuncheva, 2008].
- *Changing ensemble lineup* – replacing outdated classifiers in the ensemble, e.g., new individual models are trained on the most recent data and added to the ensemble. The ensemble pruning procedure is applied, which chooses the most valuable set of individual classifiers [Jackowski, 2014].

A comprehensive overview of classifier ensemble techniques was presented by Krawczyk et al. [Krawczyk et al., 2017]. Let us shortly characterize some popular strategies used during the experiments. *Streaming Ensemble Algorithm* (SEA) [Street and Kim, 2001] is the simple classifier ensemble with changing lineup, where the individual classifiers are trained on the successive data chunks. The base classifiers with the lowest accuracy are removed from the ensemble to keep the model up-to-date. Wang et al. proposed *Accuracy Weighted Ensembles* (AWE) [Woźniak et al., 2013] employing the weighted voting rules, where weights depend on the accuracy obtained on the testing data. Brzeziński and Stefanowski proposed *Accuracy Updated Ensemble* (AUE), extending AWE by using online classifiers and updating them according to the current distribution [Brzeziński and Stefanowski, 2011]. Woźniak et al. developed *Weighted Aging Ensemble* (WAE), which trains base classifiers on successive data chunks, and the final decision is made on weighted voting, where weights depend on accuracy and ensemble diversity. This algorithm additionally employs the decoy function to decrease the weights of outdated individuals [Woźniak et al., 2013].

2.5 Existing evaluation methodology

Because this work mainly focuses on improving classifier behavior after the *concept drift* appearance, apart from the classifier's predictive performance, we should also consider memory consumption, the time required to update the model, and time to decide. However, it should also be possible to evaluate how the model reacts to changes in the data distribution. Shaker and Hüllermeier [Shaker and Hüllermeier, 2015] presented a complete framework for evaluating the recovery rate, including the proposition of two metrics *restoration time* and *maximum performance loss*. In this framework, the notion of pure streams was introduced, i.e., streams containing only one concept. Two pure streams S_A and S_B are mixed into third stream S_C , starting with concepts only from the first stream and gradually increasing a percentage of concepts from the second stream. *Restoration time* was defined as a length of the time interval between two events - first, a performance measured on S_C drops below 95% of a S_A performance, and then the

performance on S_C rise above 95% of S_B performance. The *Maximum performance loss* is the maximum difference between S_C performance and lowest performance on either S_A or S_B . Zliobaite et al. [Zliobaite et al., 2015] proposed that evaluating the profit from the model update should consider the memory and computing resources involved in its update.

2.6 Computational and memory requirements

While designing a data stream classifier, we should also consider the computation power and memory limitations and that we usually have limited access to data labels. These data stream characteristics pose the need for other algorithms than ones previously developed for *batch learning*, where data are stored infinitely and persistently. Such learning algorithms cannot fulfill all data stream requirements, such as memory usage constraints, limited processing time, and one scan of incoming examples. However, simple incremental learning is usually insufficient, as it does not meet tight computational demands and does not tackle evolving nature of data sources [Krempel et al., 2014].

Constraints on memory and time have resulted in different windowing techniques, sampling (e.g., reservoir sampling), and other summarization approaches. Also, we have to realize that when the *concept drift* appears, data from the past may become irrelevant or even harmful for the current models, deteriorating the predictive performance of the classifiers. Thus an appropriate implementation of a forgetting mechanism (where old data instances are discarded) is crucial.

2.7 Other approaches that modify chunk size

Dynamic chunk size adaptation has been considered in the previous works. Liu et al. [Liu et al., 2017] utilize information about the occurrence of drift from drift detector. If drift occurs in the middle of the chunk, data is divided into two chunks, hence dynamic chunk size. If there is no drift inside the chunk, the whole batch is used. In the prepared chunk, the majority class is undersampled. A new classifier is trained and added to the ensemble, and older classifiers are updated. Lu et al. [Lu et al., 2020] also utilize an ensemble framework for imbalanced stream learning. In this approach, chunk size grows incrementally. Two chunks are compared based on ensembles predictions variance. An algorithm for calculating prediction variance called *subunderbagging* is introduced. Computed variance is compared using F-test. Chunk size increases if the p-value is less than a predefined threshold; otherwise, the whole ensemble is updated with the selected chunk size. The whole process repeats as long as the p-value is lower than the threshold. In both of these works, dynamic chunk size was used to handle imbalanced data streams. In contrast, we show that changing chunk size can be beneficial when handling *concept drifts* in general. Therefore, we do not focus primarily on imbalanced data.

Bifet et al. [Bifet and Gavaldà, 2007] introduced a method for handling *concept drift* with varying chunk sizes. Each incoming chunk is divided into two parts: older and new. Empirical means of data in each subchunk are compared using Hoeffding bound. If the difference between two means exceeds the threshold defined by confidence value, then data in the older window is qualified as out of date and is dropped. Later window with data for current concept grows, until next drift is detected and data is split again. This approach allows for detecting drift inside the chunk.

3 Methods

This paper presents a general framework that can be used for training any chunk-based classifier ensemble. This approach aims to reduce the restoration time, i.e., a period needed to stabilize the classification model performance after *concept drift* occurs. As we mentioned, most methods assume a fixed data chunk size, which is a parameter of these algorithms. Our proposal does not modify the core of a learning algorithm itself. Still, based on the predictive performance estimated on a given data chunk, it only indicates what data chunk size is to be taken by a given algorithm in the next step. We provide the schema of our method in Fig. 1. The intuition tells us that after the occurrence of the *concept drift*, the size of the chunk should be small to quickly train new models that will replace the models learned on the data from the previous concept in the ensemble. When the stabilization is reached, the ensemble contains base models trained on data from a new concept. At this moment, we can extend the chunk size so classifiers in the ensemble can achieve better performance and even greater stability by learning on larger portions of data from the streams because the analyzed concept is already stable.

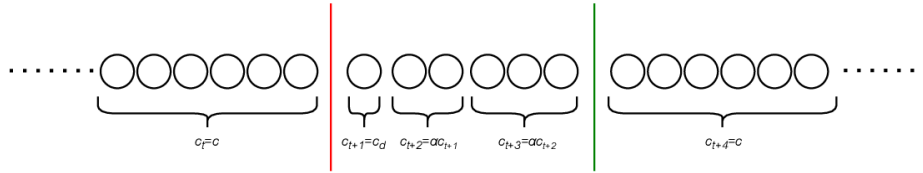


Figure 1: *Chunk-Adaptive Restoration* visualization. Red line marks the *concept drift*, green line marks the *stabilization*.

Let us present the proposed framework in detail.

3.1 Chunk-Adaptive Restoration

Starting the learning process, we sample the data from the stream with a constant chunk size c and monitor the classifier performance using a *concept drift* detector to detect changes in data distribution. When the drift occurs, we decrease the chunk size to the smaller value $c_d \ll c$, i.e., c_d is the predefined size of a batch for *concept drift*. Size of subsequent chunks after drift at given time t are computed using the following equation:

$$c_t = \min(\lfloor \alpha c_{t-1} \rfloor, c) \quad (1)$$

where $\alpha > 1$. The chunk size grows continuously with each step to reach the original value c unless the stabilization is detected. Then the chunk size is set to c immediately. Let us introduce the *Variance-based Stabilization Detection Method* (VSDM) to detect the predictive performance stabilization. First, we define the fixed-sized sliding window W containing the last K predictive performance metric values obtained for the most recent chunks. We also introduce the stabilization threshold ϵ_s . The stabilization is detected when the following condition is met:

$$\text{Var}(W) < \epsilon_s \quad (2)$$

where $Var(W)$ is a variance of scores obtained for the last K chunks. Sample data stream with detected drift and stabilization is presented in Fig. 2. The primary assumption of the proposed method is a faster model adaptation caused by the increased number of updates after a *concept drift*. This strategy allows for using the larger chunk sizes when the data is not changing. It also reduces the computational costs of retraining models. Alg. 1 present the whole procedure. Our method works with existing models for online learning. For this reason, we argue that the approach proposed in this paper is easier to deploy in practice.

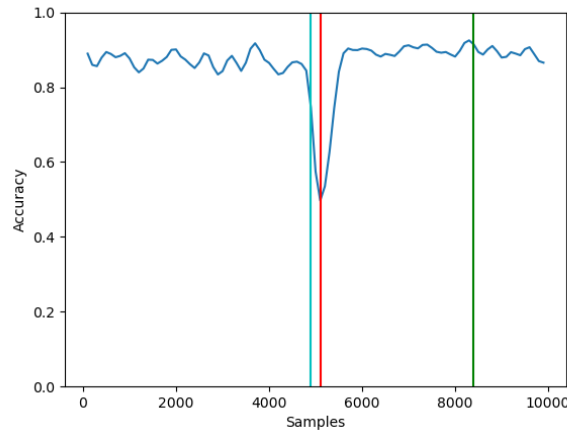


Figure 2: Exemplary accuracy for data stream with abrupt concept. Red line denotes drift detection, green stabilization detection, and blue beginning of a real drift.

3.2 Memory and time complexity

Our method only impacts the size of the chunk. All other factors like the number of features or classifiers in the ensemble are the same as in the basic approach. For this reason, we will focus here only on the impact of chunk size on memory and time complexity. With memory complexity, our method could impact only the size of buffers for storing samples from a stream. When no drift is detected, the standard chunk size is used. This dictates the required size of buffers for storing samples. For this reason, memory complexity for storing samples is $O(c)$.

CAR works the same way as a base method when no drift is detected and the data stream is stable. Therefore, in this case, the time complexity is the same as in the base method. When drift is detected sizes of subsequent chunks are changed. Time complexity depends on model complexity $g(N)$, where N is the number of learning examples provided to model to train on. For simplicity, we assume that $g(N)$ represents both ensemble and base model complexity. With this assumptions time complexity of base model (when CAR is not enabled) is: $O(g(c))$. When CAR is enabled, and *concept drift* is detected, the chunk size is changed to c_d . Each consecutive chunk at time t has size

Algorithm 1 Chunk-Adaptive Restoration algorithm**Input:** m - model S - data stream dd - drift detector sd - stabilization detector n - number of chunks t - chunk index c - base chunk size c_d - base drift chunk size c_t - t th chunk size $test()$ - procedure that tests model with a chunk and returns the predictive performance metric (ppm) $train()$ - procedure that trains model with a chunk $change_detected()$ - procedure that informs about drift occurrence with the drift detector and the last score $stabilization_detected()$ - procedure that detects stabilization with the stabilization detector and the stabilization window

```

1: for  $t = 1$  to  $n$  do
2:    $ppm \leftarrow test(m, S(t))$ 
3:   if  $stabilization\_detected(sd, ppm)$  then
4:      $c_t \leftarrow c$ 
5:   else
6:      $c_t \leftarrow \min(\lfloor \alpha c_{t-1} \rfloor, c)$ 
7:   end if
8:   if  $change\_detected(dd, ppm)$  then
9:      $c_t \leftarrow c_d$ 
10:  end if
11:   $train(m, S(t))$ 
12: end for

```

$c_t = \alpha^t c_d$, with $t = 0$ directly after the drift was detected. Chunk size grows until stabilization is detected or current chunk size is restored to original size c . For simplicity, we skip the case when stabilization is detected. With this assumption, we write condition for restoring the original chunk size:

$$\alpha^{t_s} c_d = c \quad (3)$$

where t_s is the time when the chunk size is restored to its original value. From this equation, we obtain t_s directly:

$$t_s = \log_{\alpha} \frac{c}{c_d} \quad (4)$$

The number of operations required by CAR after *concept drift* was detected is

$$\sum_{t=0}^{t_s} g(\alpha^t c_d) \quad (5)$$

Using big-O notation:

$$O\left(\sum_{t=0}^{t_s} g(\alpha^t c_d)\right) = O(g(\alpha^{t_s} c_d)) = O\left(g\left(\frac{c}{c_d}\right)\right) = O(g(c)) \quad (6)$$

Therefore CAR time complexity depends only on chunk size and computational complexity of used models.

3.3 Sample Restoration

Restoration time cannot be directly utilized in this work, as we do not have access to pure streams with separate concepts. For this reason, we introduce a new Sample Restoration (SR) metric to evaluate the *Chunk-Adaptive Restoration* performance compared to standard methods used for learning models on data streams with *concept drift*. We assume that there is a sequence of N chunks between two stabilization points. Each element of such a sequence is determined by the chunk size c_t and the achieved model's accuracy acc_t . Let us define the index of the minimum accuracy as:

$$t_{min} = \underset{t \in [0, N)}{\operatorname{argmin}} acc_t \quad (7)$$

and the restoration threshold is given by the following formula:

$$r = p \times \max_{t \in [t_{min}, N)} acc_t \quad (8)$$

where $p \in (0, 1)$ is the percentage of the performance that has to be restored, and the multiplier is the maximum accuracy score of our model after the point when it achieved its minimum score. Finally, we look for the lowest index t_r after which the model exceeds the assumed restoration threshold:

$$t_r = \inf_{t \in [t_{min}, N)} \{t : acc_t \geq r\} \quad (9)$$

Sample Restoration is computed as the sum of chunk sizes from the *concept drift*'s beginning to the t_r :

$$SR(p) = \sum_{t=0}^{t_r} c_t \quad (10)$$

In general, SR is the number of samples needed to obtain the p percent of the maximum performance achieved on the subsequent task.

4 Experiment

Chunk-Adaptive Restoration is a method designed to reduce the number of samples used to restore the model's performance during the *concept drift*. We expect to significantly

reduce the *Sample Restoration* for each trained model depending on the chunk size adaptation level. The experimental study was formulated to answer the following research questions:

RQ1: How do different chunk sizes impact predictive performance?

RQ2: How does the *Chunk-Adaptive Restoration* influence the learning process?

RQ3: How many samples can be saved during the restoration phase?

RQ4: How do different classifier ensemble models behave with the application of *Chunk-Adaptive Restoration*?

RQ5: How robust to noise the *Chunk-Adaptive Restoration* is?

4.1 Experiment setup

Data streams. Experiments were carried out using both synthetic and real datasets. Stream-learn library [Ksieniewicz and Zyblewski, 2020] was employed to generate the synthetic data containing three types of *concept drift*: abrupt, gradual, and increment, all generated with the recurring or unique concepts. We tested parameters such as chunk sizes and the stream length for each type of *concept drift*. All streams were generated with 5 *concept drifts*, 2 classes, 20 input features, of which 2 were informative, and 2 were redundant. In the case of incremental and gradual drifts concept, sigmoid spacing was set to 5. Apart from the synthetic ones, we employed the Usenet [Katakis et al., 2010] and Insects [Souza et al., 2020] data streams. Unfortunately, the original Usenet dataset contains a small number of samples, so two selected concepts were repeated to create a recurring-drifted data stream. Each chunk of the Insects data stream was randomly oversampled because of the significant imbalance ratio. Tab. 1 contains detailed description of all utilized data streams.

Drift detector. The Fast Hoeffding Drift Detection Method [Blanco et al., 2015] was employed as a *concept drift* detector. We used implementation available on the public repository [w4k2, 2021]. The size of a window in FHDDM was equal to 1000, and the error probability allowed $\delta = 10^{-6}$.

Classifier ensembles. Three models classifier ensembles dedicated to data stream classification were chosen for comparison:

- Weighted Aging Classifier (WAE) [Woźniak et al., 2013]
- Accuracy Weighted Ensemble (AWE) [Wang et al., 2003],
- Streaming Ensemble Algorithm (SEA) [Street and Kim, 2001],

All ensembles contained 10 base classifiers.

Experimental protocol. In our experiments, we apply the models mentioned above to selected data streams with *concept drift*. We measure Sample Restoration. These results are reported as a baseline. Next, we apply Chunk-Adaptive Restoration and repeat experiments to establish the proposed model's influence on the ability to handle *concept drift* quickly. As the experiments were conducted with the balanced data, the accuracy was used as the only indicator of the model's performance. As the experimental protocol, *Test-Then-Train* was employed [Bifet et al., 2010].

#	Source	Drift type	Base chunk size c	#samples
1	stream-learn	abrupt recurring	500	300000
2	stream-learn	abrupt recurring	1000	150000
3	stream-learn	abrupt recurring	10000	60000
4	stream-learn	abrupt recurring	500	250000
5	stream-learn	abrupt nonrecurring	500	300000
6	stream-learn	abrupt nonrecurring	1000	150000
7	stream-learn	abrupt nonrecurring	10000	60000
8	stream-learn	abrupt nonrecurring	500	250000
9	stream-learn	gradual recurring	500	300000
10	stream-learn	gradual recurring	1000	150000
11	stream-learn	gradual recurring	10000	60000
12	stream-learn	gradual recurring	500	250000
13	stream-learn	gradual nonrecurring	500	300000
14	stream-learn	gradual nonrecurring	1000	150000
15	stream-learn	gradual nonrecurring	10000	60000
16	stream-learn	gradual nonrecurring	500	250000
17	stream-learn	incremental recurring	500	300000
18	stream-learn	incremental recurring	1000	150000
19	stream-learn	incremental recurring	10000	60000
20	stream-learn	incremental recurring	500	250000
21	stream-learn	incremental nonrecurring	500	300000
22	stream-learn	incremental nonrecurring	1000	150000
23	stream-learn	incremental nonrecurring	10000	60000
24	stream-learn	incremental nonrecurring	500	250000
25	usenet	abrupt recurring	1000	120000
26	insects-abrupt-imbalanced	abrupt nonrecurring	1000	355275
27	insects-gradual-imbalanced	gradual nonrecurring	1000	143323

Table 1: Data streams used for experiments.

Statistical analysis. Because Sample Restoration can be computed for each drift and *concept drift* can occur multiple times, we report average Sample Restoration for each stream with standard deviation. To assess the statistical significance of the results, we used a one-sided Wilcoxon signed-rank test in a direct comparison between the models with the 95% confidence level.

Reproducibility. To enable independent reproduction of our experiments, we provide a GitHub repository with code ¹. This repository also contains detailed results of all experiments. Stream-learn [Ksieniewicz and Zyblewski, 2020] implementation of the ensemble models was utilized with the Gaussian Naïve Bayes and CART as base classifiers taken from the Scikit Learn library [Pedregosa et al., 2011]. Detailed information about used packages is provided in the YAML file with a specification of the Anaconda

¹ <https://github.com/w4k2/chunk-adaptive-restoration>

environment.

4.2 Impact of chunk size on performance

In our first experiment, we examine the impact of the chunk size on the model performance and general capability for handling data with *concept drift*. To evaluate these properties, we train the AWE model on a synthetic data stream with different chunk sizes. The stream consists of 20 features, 2 classes, and it contains only 1 abrupt drift. Results are presented in Fig. 3. As expected, chunk size has an impact on the maximal accuracy that the model can achieve. It is especially visible before a drift, where models with larger chunks obtain the best accuracy. Also, with larger chunks variance of accuracy is lower. In ensemble-based approaches, a base classifier is trained on a single chunk. A larger chunk means that more data is available to the underlying model. Therefore it allows for the training of a more accurate model. Interestingly we can see that for all chunk sizes, performance is restored roughly at the same time. Regardless of the chunk size, a similar number of updates is required to bring back the model performance. Please keep in mind that the x-axis in Fig. 3 is the number of chunks. It means that models trained on larger chunks require a larger number of learning examples to restore accuracy.

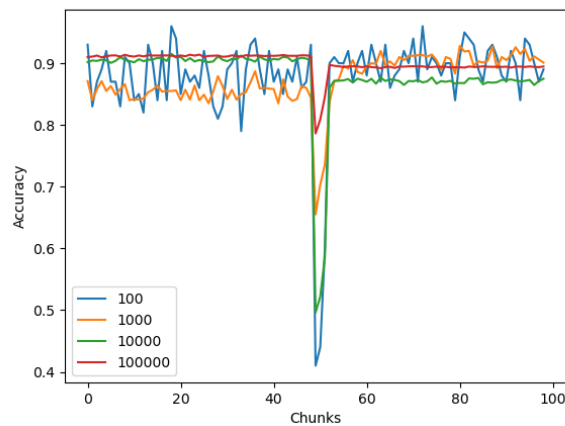


Figure 3: Impact of chunk size on obtained accuracy.

These results give the rationale behind our method. When drift is detected, we change chunk size to decrease the consumption of learning examples required for restoring accuracy. Next, we gradually increase chunk size to improve the maximum possible performance when the model recovers from drift. It allows for a quick reaction to drift and does not limit the model's maximum performance. In principle, not all models are compatible with changing chunk size. Also, batch size cannot be decreased indefinitely. Minimal chunk size should be determined case by case, dependent on the base learner used in an ensemble or used model in general. Later in our experiments, we use chunk sizes of 500, 1000, and 10000 to obtain a reliable estimate of how our method will perform in different settings.

4.3 Hyperparameter tuning

After chunk size was selected, we fine-tuned other hyperparameters, and then we proceeded to further experiments. Firstly set two values manually, based on our observations. First is α (i.e., constant that determines how fast chunk size grows after drift was detected) equal to 1.1. Second is drift chunk size equal to 30, as it is a typical window length in drift detectors.

Next, we find the best stabilization window size and the threshold. We conduct grid search with windows size values 30, 50, 100, and stabilization thresholds 0.1, 0.01, 0.001, 0.0001. For experiments, we use synthetic data streams 1-24 from the Tab. 1. Used data streams have different random number generator seeds in this and later experiments. Results were collected for WAE, AWE, SEA ensembles with Naïve Bayes base model. We use Sample Restoration 0.8 as a performance indicator. For each set of parameters, Sample Restoration was averaged over all streams used to obtain one value. Results are provided in the Tab. 2.

stabilization thresholds	drift chunk size		
	30	50	100
0.1	59210.11	59210.11	59210.11
0.01	58489.47	58675.99	58709.98
0.001	55328.20	55363.95	57669.70
0.0001	52846.04	55962.58	62398.56

Table 2: Sample Restoration 0.8 for various hyperparameter setting. Lower is better.

From provided data, we can conclude that the smaller the drift chunk size, the lower the SR is. This observation is in line with intuition about our method. Smaller drift chunk size provides a larger benefit during drift compared to normal chunk size. The same dependency can be observed for the stabilization threshold. Intuitively, a lower threshold means that stabilization is harder to reach. We argue that this can be beneficial in some cases when working with gradual or incremental drift. In this scenario, if stabilization is reached too fast, then chunk size is immediately brought back to the standard size, and there is no benefit from a smaller chunk size at all. Lowering the stabilization threshold could help in these cases. In later experiments, we use the stabilization window size equal to 30 and the variance stabilization threshold equal to 0.0001.

4.4 Impact on *concept drift* handling capability

In this part of the experiments, we compare the performance of the proposed method to baseline. Results were collected following the experimental protocol described in the previous sections. To save space, we do not provide results for all models and streams. Instead, we plot accuracy achieved by models on selected data streams. These results are presented in Fig. 4, 5, 6, and 7. All learning curves were smoothed using a 1D Gaussian filter with $\sigma = 1$.

From provided plots, we can deduce that the largest gains from employing the CAR method can be observed for an abrupt data stream. In streams with gradual and incremental drifts, there are fewer or none sudden drops of accuracy that the model can quickly react to. For this reason, the CAR method does not provide a large benefit

with this kind of *concept drifts*. During a more detailed analysis of obtained results, we observed that the stabilization for gradual and incremental drifts is hard to detect. Many false positives usually cause an early return to the original chunk size, influencing the performance achieved on those two types of drifts. FHDDM caused another problem regarding the early detection of the gradual and incremental *concept drifts*. Usually, this is a desired feature. In our method, early drift detection initiates the chunk size change when two data concepts are still overlapping during stream processing. As the transition between two concepts takes much time, when one concept starts to dominate, the chunk size could be restored to its original value too early, affecting the achieved results.

We also observe larger gains from applying CAR on streams with bigger chunk sizes. To illustrate please compare results from Fig. 4 to Fig. 5. One possible explanation behind this trend is that gains obtained from employing CAR are proportional to the difference in size between the base and drift chunk size. In our experiments, drift chunk size was equal to 30 for all streams and models. This explanation is also in line with the results of hyperparameter experiments provided in the Tab. 2.

We conclude this section by providing a statistical analysis of our results. Tab. 3 shows the results of the Wilcoxon test for Naïve Bayes and CART base models. We state meaningful differences in the Sample Restoration between the baseline and the CAR method for all models.

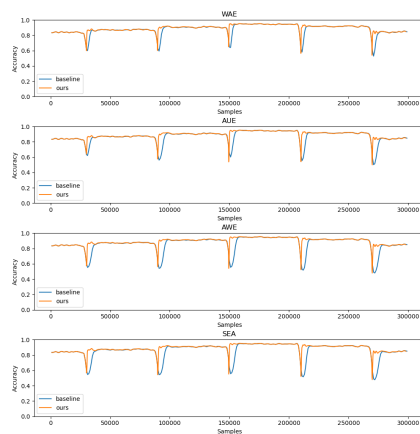


Figure 4: Accuracy for stream-learn data stream (1).

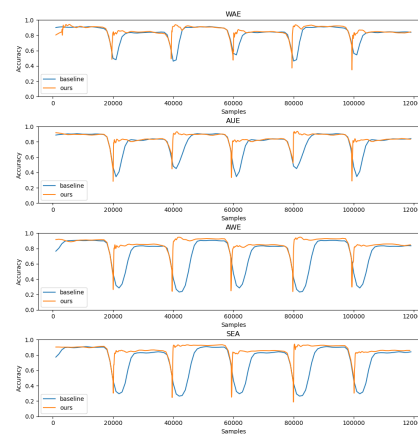


Figure 5: Accuracy for Usenet dataset (25).

4.5 Impact of noise on the CAR effectiveness

Real-world data often contain noise in labeling. For this reason, we evaluate if the proposed method can be used for data with varying amounts of noise in labels. We generate a synthetic data stream with two classes, base chunk size 1000, drift chunk size 100, and single, abrupt *concept drift*. We randomly select a predefined fraction of samples in each chunk and flip labels for selected learning examples. Next, we measure the accuracy of the AUE model with Gaussian Naïve Bayes base model on a generated

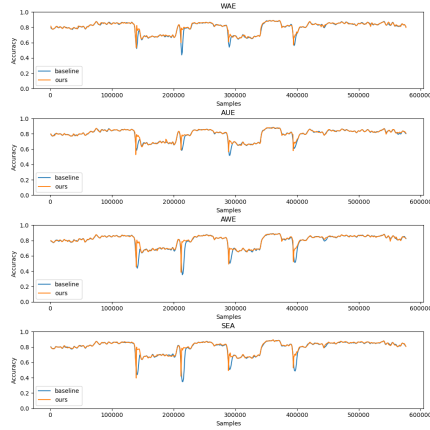


Figure 6: Accuracy for abrupt Insects dataset (26).

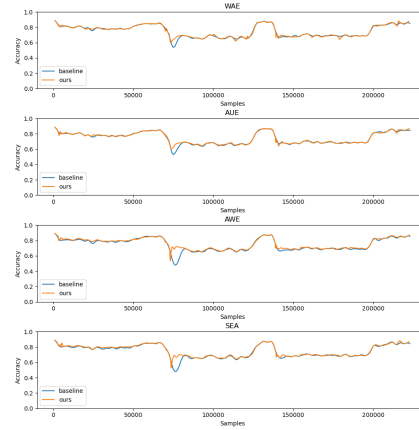


Figure 7: Accuracy for gradual Insects dataset (27).

Naïve Bayes						
model	SR(0.9)		SR(0.8)		SR(0.7)	
name	Statistic	p-value	Statistic	p-value	Statistic	p-value
WAE	40.0	0.0006	30.0	0.0002	45.0	0.0009
AWE	22.0	9.675e-05	26.0	0.0001	36.0	0.0004
SEA	0.0	1.821e-05	23.0	0.0001	1.0	1.389e-05
Cart						
model	SR(0.9)		SR(0.8)		SR(0.7)	
name	Statistic	p-value	Statistic	p-value	Statistic	p-value
WAE	14.0	6.450e-05	54.0	0.003	55.0	0.003
AWE	0.0	1.229e-05	6.0	2.543e-05	21.0	0.0001
SEA	23.0	0.0001	43.0	0.001	42.0	0.001

Table 3: Wilcoxon test results

dataset with noise levels 0, 0.1, 0.2, 0.3, and 0.4. Results are presented in Fig. 8. We note that for low noise levels, i.e., up to 0.3, restoration time is shorter. With a larger amount of noise, there is no sudden drop in accuracy. Therefore CAR has no impact on the speed of reaction to drift.

It should be noted that results for CAR with noise levels 0.2, 0.3, and 0.4 were generated with the stabilization detector turned off. With a higher amount of noise, stabilization was detected very fast. Therefore chunk size was quickly set to base value. In this case, there was no benefit of applying CAR. This indicates that the stabilization method should be refined to handle noisy data well.

4.6 Lessons learned

Firstly we evaluated the impact of chunk size on the process of learning in the data stream with single *concept drift*. We learn that models with larger chunk size can obtain larger

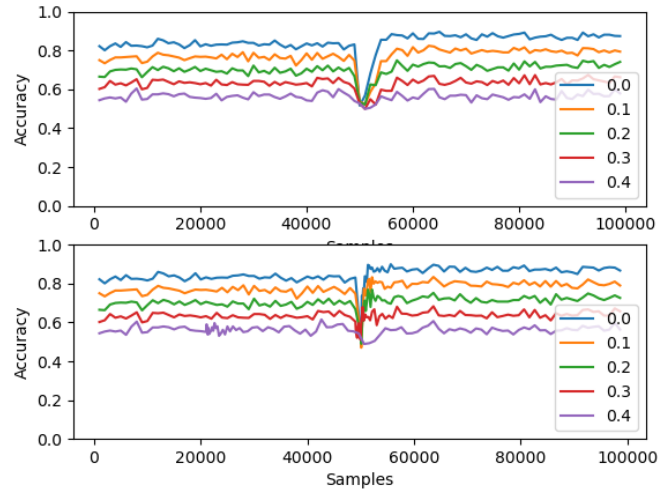


Figure 8: Impact of noise in labels on proposed method effectiveness. (Upper) baseline accuracy for synthetic data stream with different noise level added to labels. (Lower) CAR accuracy for the same synthetic data stream. In case of Noise levels 0.2, 0.3, and 0.4 stabilization detector was turned off.

maximum accuracy, but the required number of updates to restore accuracy is similar regardless of chunk size (RQ1 answered). The main goal of introducing the *Chunk-Adaptive Restoration* was to prove its advantages in controlling the number of samples during the restoration period while dealing with abrupt *concept drift*. The statistical tests have shown a significant benefit of employing it in different stream learning scenarios (RQ2 answered). The method's highest gains were observed when the large original chunk size was used. There are fewer model updates with a bigger chunk size, resulting in a delay of reaction to *concept drift*.

The number of samples that can be saved depends on the drift type and the original chunk size. When dealing with abrupt drift, the sample restoration time can be around 50% better than the baseline (RQ3 answered). We noticed that CAR minimized restoration time for each of the analyzed classifier ensemble methods and achieved better average predictive performance. It is worth noting that the simpler the algorithm, the greater the profit from using CAR. The most considerable profit was observed for SEA and AWE, while in the case of WAE, sometimes the native version outperformed CAR for the Average Sample Restoration metric (RQ4 answered). When a small amount of noise is present in labels, CAR can still be useful, however in some cases stabilization detector should not be used. With a larger amount of noise, there is no gain from using the proposed method (RQ5 answered).

5 Conclusion

The work focused on the *Chunk-Adaptive Restoration* framework, which is dedicated to chunk-based data stream classifiers enabling better recovery from *concept drifts*. We

proposed new methods for stabilization detection and chunk size adaptation. It is worth emphasizing that any block-based data stream classifier can use CAR, especially for the tasks where *concept drift* may appear. We evaluated the developed algorithms based on extensive experimental studies using real and synthetic data streams. Obtained results show a significant difference between the predictive performance of the baseline models and models employing CAR. *Chunk-Adaptive Restoration* is strongly recommended for abrupt *concept drift* scenarios because it significantly can reduce model downtime. The performance gain is not visible for other types of *concept drift*, but it still achieves acceptable results. The future works may focus on:

- Improving the *Chunk-Adaptive Restoration* behavior for gradual and incremental *concept drifts*.
- Adapting the *Chunk-Adaptive Restoration* to the case of limited access to labels using a semi-supervised and active learning approach.
- Proposing a more flexible method of changing data chunk size, e.g., based on the model stability assessment.
- Adapting the proposed method to imbalanced data stream classification task, where changing the data chunk size may be correlated with the intensity of data preprocessing (e.g., the intensity of data oversampling).
- Improve stabilization method to better handle data streams with label and attribute noises.

Acknowledgements

This work is supported by the CEUS-UNISONO programme, with funding from the National Science Centre, Poland under grant agreement No. 2020/02/Y/ST6/00037.

References

- [Baena-Garcia et al., 2006] Baena-Garcia, M., del Campo-Ávila, J., Fidalgo, R., Bifet, A., Gavaldà, R., and Morales-Bueno, R. (2006). Early drift detection method. In *Fourth international workshop on knowledge discovery from data streams*, volume 6, pages 77–86.
- [Bifet et al., 2018] Bifet, A., Gavaldà, R., Holmes, G., and Pfahringer, B. (2018). *Machine Learning for Data Streams: With Practical Examples in MOA*. The MIT Press.
- [Bifet and Gavaldà, 2007] Bifet, A. and Gavaldà, R. (2007). Learning from time-changing data with adaptive windowing. volume 7.
- [Bifet et al., 2010] Bifet, A., Holmes, G., Kirkby, R., and Pfahringer, B. (2010). Moa: Massive online analysis. *J. Mach. Learn. Res.*, 11:1601–1604.
- [Bifet et al., 2009] Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., and Gavaldà, R. (2009). New ensemble methods for evolving data streams. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '09*, pages 139–148, New York, NY, USA. ACM.
- [Blanco et al., 2015] Blanco, I. I. F., del Campo-Avila, J., Ramos-Jimenez, G., Bueno, R. M., Diaz, A. A. O., and Mota, Y. C. (2015). Online and non-parametric drift detection methods based on hoeffding's bounds. *IEEE Trans. Knowl. Data Eng.*, 27(3):810–823.

- [Brzeziński and Stefanowski, 2011] Brzeziński, D. and Stefanowski, J. (2011). Accuracy updated ensemble for data streams with concept drift. In Corchado, E., Kurzyński, M., and Woźniak, M., editors, *Hybrid Artificial Intelligent Systems*, pages 155–163, Berlin, Heidelberg. Springer Berlin Heidelberg.
- [Duda et al., 2001] Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley, New York, 2. edition.
- [Gama et al., 2004] Gama, J., Medas, P., Castillo, G., and Rodrigues, P. (2004). Learning with drift detection. In *In SBIA Brazilian Symposium on Artificial Intelligence*, pages 286–295. Springer Verlag.
- [Gustafsson, 2000] Gustafsson, F. (2000). *Adaptive Filtering and Change Detection*. Wiley.
- [Harvey et al., 2021] Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S. J., Robertson, D. L., and Consortium, C.-G. U. C.-U. (2021). Sars-cov-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, 19(7):409–424.
- [Jackowski, 2014] Jackowski, K. (2014). Fixed-size ensemble classifier system evolutionarily adapted to a recurring context with an unlimited pool of classifiers. *Pattern Analysis and Applications*, 17(4):709–724.
- [Joao et al., 2014] Joao, G., Žliobaite, Albert, B., Mykola, P., and Abdelhamid, B. (2014). A survey on concept drift adaptation. 46(4).
- [Junsawang et al., 2019] Junsawang, P., Phimoltares, S., and Lursinsap, C. (2019). Streaming chunk incremental learning for class-wise data stream classification with fast learning speed and low structural complexity. *PLoS one*, 14(9):e0220624.
- [Katakis et al., 2010] Katakis, I., Tsoumakas, G., and Vlahavas, I. (2010). Tracking recurring contexts using ensemble classifiers: An application to email filtering. *Knowledge and Information Systems*, 22:371–391.
- [Krawczyk et al., 2017] Krawczyk, B. et al. (2017). Ensemble learning for data stream analysis: A survey. *Inf. Fusion*, 37:132 – 156.
- [Krempel et al., 2014] Krempel, G., Žliobaite, I., Brzeziński, D., Hüllermeier, E., Last, M., Lemaire, V., Noack, T., Shaker, A., Sievi, S., Spiliopoulou, M., and Stefanowski, J. (2014). Open challenges for data stream mining research. *SIGKDD Explor. Newsl.*, 16(1):1–10.
- [Ksieniewicz and Zybiewski, 2020] Ksieniewicz, P. and Zybiewski, P. (2020). stream-learn—open-source python library for difficult data stream batch analysis. *arXiv preprint arXiv:2001.11077*.
- [Kuncheva, 2004] Kuncheva, L. I. (2004). Classifier ensembles for changing environments. In Roli, F., Kittler, J., and Windeatt, T., editors, *Multiple Classifier Systems, 5th International Workshop, MCS 2004, Cagliari, Italy, June 9-11, 2004, Proceedings*, volume 3077 of *Lecture Notes in Computer Science*, pages 1–15. Springer.
- [Lazarescu et al., 2004] Lazarescu, M. M., Venkatesh, S., and Bui, H. H. (2004). Using multiple windows to track concept drift. *Intell. Data Anal.*, 8(1):29–59.
- [Liu et al., 2017] Liu, N., Zhu, W., Liao, B., and Ren, S. (2017). Weighted ensemble with dynamical chunk size for imbalanced data streams in nonstationary environment.
- [Lu et al., 2020] Lu, Y., Cheung, Y.-M., and Yan Tang, Y. (2020). Adaptive chunk-based dynamic weighted majority for imbalanced data streams with concept drift. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):2764–2778.
- [Oza and Tumer, 2008] Oza, N. C. and Tumer, K. (2008). Classifier ensembles: Select real-world applications. *Inf. Fusion*, 9(1):4–20.

- [Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- [Raudys, 2014] Raudys, S. (2014). *Statistical and Neural Classifiers: An Integrated Approach to Design*. Springer Publishing Company, Incorporated.
- [Rodríguez and Kuncheva, 2008] Rodríguez, J. J. and Kuncheva, L. I. (2008). Combining online classification approaches for changing environments. In *Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, SSPR & SPR '08*, pages 520–529, Berlin, Heidelberg. Springer-Verlag.
- [Shaker and Hüllermeier, 2015] Shaker, A. and Hüllermeier, E. (2015). Recovery analysis for adaptive learning from non-stationary data streams: Experimental design and case study. *Neurocomputing*, 150:250–264.
- [Sobolewski and Woźniak, 2013] Sobolewski, P. and Woźniak, M. (2013). Concept drift detection and model selection with simulated recurrence and ensembles of statistical detectors. *Journal of Universal Computer Science*, 19(4):462–483.
- [Souza et al., 2020] Souza, V. M. A., Reis, D. M., Maletzke, A. G., and Batista, G. E. A. P. A. (2020). Challenges in benchmarking stream learning algorithms with real-world data. *Data Mining and Knowledge Discovery*, pages 1–54.
- [Street and Kim, 2001] Street, N. and Kim, Y. (2001). A streaming ensemble algorithm (sea) for large-scale classification. pages 377–382.
- [w4k2, 2021] w4k2 (2021). Chunk adaptive restoration. <https://github.com/w4k2/chunk-adaptive-restoration>.
- [Wang et al., 2003] Wang, H., Fan, W., Yu, P. S., and Han, J. (2003). Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03*, page 226–235, New York, NY, USA. Association for Computing Machinery.
- [Widmer and Kubat, 1993] Widmer, G. and Kubat, M. (1993). Effective learning in dynamic environments by explicit context tracking. In Brazdil, P., editor, *Machine Learning: ECML-93*, volume 667 of *Lecture Notes in Computer Science*, pages 227–243. Springer Berlin Heidelberg.
- [Widmer and Kubat, 1996] Widmer, G. and Kubat, M. (1996). Learning in the presence of concept drift and hidden context. *Machine Learning*, 23:69–101.
- [Woźniak et al., 2013] Woźniak, M., Kasprzak, A., and Cal, P. (2013). Weighted aging classifier ensemble for the incremental drifted data streams. In Larsen, H. L., Martin-Bautista, M. J., Vila, M. A., Andreasen, T., and Christiansen, H., editors, *Flexible Query Answering Systems*, pages 579–588, Berlin, Heidelberg. Springer Berlin Heidelberg.
- [Zliobaite et al., 2015] Zliobaite, I., Budka, M., and Stahl, F. T. (2015). Towards cost-sensitive adaptation: When is it worth updating your predictive model? *Neurocomputing*, 150:240–249.