ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Data stream generation through real concept's interpolation

Joanna Komorniczak and Pawel Ksieniewicz

Department of Systems and Computer Networks
Wroclaw University of Science and Technology
Wroclaw, Poland

Abstract. Among the recently published works in the field of data stream analysis – both in the context of *classification* task and *concept drift detection* – the deficit of real-world data streams is a recurring problem. This article proposes a method for generating data streams with given parameters based on real-world static data. The method uses one-dimensional interpolation to generate *sudden* or *incremental* concept drifts. The generated streams were subjected to an exemplary analysis in the *concept drift detection* task with a detector ensemble. The method can potentially contribute to the development of methods focused on data stream processing.

1 Introduction

Data stream processing has been a frequent subject of research in recent years [1]. A distinctive feature of streams is a potentially infinite inflow of data, which requires reliable methods [2]. The most often considered difficulty occurring in data streams is *concept drifts*, which negative impact on classification models can be reduced with drift detectors or adaptive measures [3].

Often, the available methods are based on the analysis of synthetic data streams [4, 5]. A significant advantage of this type of stream acquisition is the ability to specify stream parameters to obtain data with specific characteristics, such as the number of features, the number of drifts, and the length of the stream. Additionally, synthetic data is easy to reproduce and can be explicitly regenerated instead of stored [6], which is an important issue in any incremental learning scenario. Despite the great convenience resulting from synthetic data, we should avoid performing the model's evaluation on the artificial problems as it may favor one of the models [7].

The best approach is to evaluate methods on real-world data. Unfortunately, the availability of real-world data streams characterized by concept drifts with indications of drift moments is limited [4, 8]. The ones available are either too simple in terms of classification task, drift types may vary over stream course, or – as mentioned before – the moments of drifts are not designated. Often, the classification qualities over the entire stream processing are used as a measure of the effectiveness of drift detection. This measure may not provide enough information about the quality of detection. It is worth comparing the moments of drift detection with the moments of actual drift [9]. By increasing the availability of streams containing drift ground-truth, we can contribute to the development of effective drift detection methods.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

This work proposes a method for generating data streams with concept drifts using real-world static datasets. The method allows users to specify the characteristics of the generated stream, including the number of drifts and their type, which allows for extensive research on streams with different parameters. The implementation is available in the GitHub repository [1] as well as a set of sample data streams in the form of a ZIP archive.

2 Method

This Section intends to describe the operation of the proposed stream generator.

The generator was implemented in *Python* programming language using the *Scipy* and *numpy* libraries. The method takes a static dataset as an input parameter. The user can specify the length of the stream – the total number of samples, the final number of features, and the number of drifts. For the generation of drifts, the method uses one-dimensional interpolation. The choice of the interpolation method enables the generation of a given type of drift – *nearest* which corresponds to *sudden* and *linear* or *cubic*, which resembles *incremental* drift [10].

The generator then performs random over– or undersampling – depending on whether the desired stream length is greater or smaller than the original data. The prior probability of the input data is maintained. Additionally, after resampling, the data is randomly shuffled. Depending on the number of drifts, a certain number of central points of the concept's occurrence is selected. The distances between the concepts' central points are equal, allowing the calculation of drift occurrence moments.

Base projections are drawn from the normal distribution. A matrix with dimensions corresponding to the number of base features per target features is drawn for each existing concept. The purpose of the projections is to enable the generation of an output stream with a specified feature number and to perform concept drift generation. Sample one-dimensional projections are presented in Figure 1.

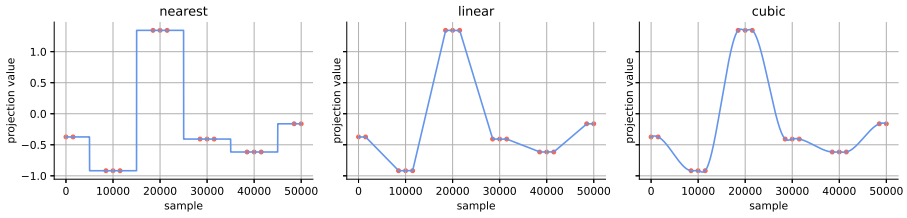

Fig. 1: Projection values and concept basepoints of different kinds of drifts – *nearest*, *linear* and *cubic* for stream with 5 drifts

In the case of incremental drifts, we should ensure that the concept changes

[1] `https://github.com/w4k2/ip_stream_generator`

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

are not constant – there is a need to stabilize a given concept. For this purpose, auxiliary points are selected. The distance to them from the concept center point is specified by the parameter and described by *stabilization factor*, which specifies the fraction of the single concept duration in which additional points will be sampled. Based on the interpolated central points of the concept, a continuous feature projection is generated over the entire stream length. The projection is then multiplied by the original data, forming a result stream. The generation process is also described in the Algorithm 1.

Algorithm 1 Pseudocode of the *Stream Generator*

Input:
$\mathcal{X}$, y – original static data, s – stream length in samples, f – feature number of output stream, $\mathcal{D}n$ – number of drifts in output stream, $\mathcal{I}t$ – interpolation type, S_f – stabilization factor
Output:
$DS : \{X_s, y\}$ – output stream

▷ oversample / undersample
$\mathcal{X} \leftarrow \varnothing$
$y \leftarrow \varnothing$
for each class C **in** y **do**
  $IR \leftarrow \frac{|y=C|}{|y|}$
  $i_C \leftarrow IR \cdot s$ random indexes of samples with class C
  $\mathcal{X} \leftarrow \mathcal{X} \cup \mathcal{X}[i_C]$
  $y \leftarrow y \cup y[i_C]$
end for
shuffle $\mathcal{X}, y$

▷ find concept basepoints
$c\_b \leftarrow$ evenly distributed $\mathcal{D}n + 1$ points
$projections \leftarrow$ array of random values from *normal distribution* of size ($|c\_b|$ x original feature number x f)

▷ add auxiliary concept basepoints
for each c **in** c_b **do**
  $a_{-1} \leftarrow c - \frac{s}{\mathcal{D}n} \cdot S_f$
  $a_0 \leftarrow c$
  $a_1 \leftarrow c + \frac{s}{\mathcal{D}n} \cdot S_f$
  replace c in c_b by $[a_{-1}, a_0, a_1]$
  replace each p from $projections$ with $[p, p, p]$
end for
Remove first and last element of c_b, $projections$

▷ generate continous projections
initialize array $continous\_projections$
for each s_f **in range** f **do**
  for o_f **in range** original features **do**
    define interpoation $\mathcal{I}t$ on $projections$ for current s_f, o_f in c_b
    $cp \leftarrow$ interpolated values of length s
    add cp to $continous\_projections$
  end for
end for

▷ generate stream features
$\mathcal{X}s \leftarrow continous\_projections \cdot \mathcal{X}$
return $DS : \{X_s, y\}$

3 Stream visualization

This Section describes a use case of the proposed generator and presents generated data streams with concept drift.

Cubic and *nearest* drift types were selected for presentation. Figure 2 shows the average value of the features in each chunk. Thanks to that, changes in the

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

feature's values over the stream course are visible. It is noticeable that *cubic* drift type offers more fluid changes compared to *nearest*. In the case of *cubic* drift, the stabilization of a concept is also visible.

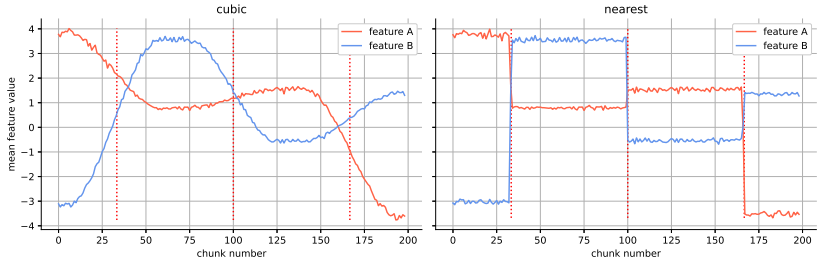

Fig. 2: Generated streams with three concept drifts of type *cubic* and *nearest*

Additionally, Figure 3 presents a feature distribution over the stream course. Each sub-plot presents a distribution of single batch samples for different concepts of generated streams. The color of a sample corresponds to an instance class. Not only the feature values are changing – which is noticeable by labels on x and y axis – but also the posterior distribution of stream features.

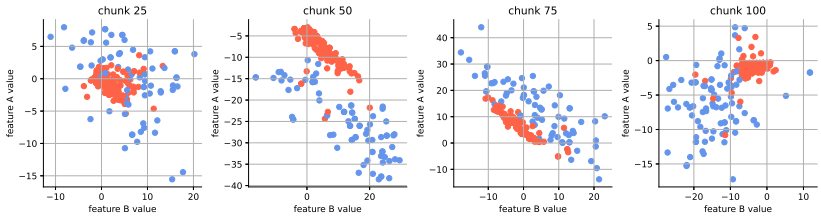

Fig. 3: Individual concepts of stream generated using *wisconsin* dataset

4 Generated stream employment

Using the proposed method and publicly available datasets from the *keel* repository, streams of *nearest* and *cubic* drift types were generated. A concept drift detection ensemble and an exemplary experiment were prepared to ensure the proposed stream generation method is competent in generating streams with research potential. The ensemble was developed using the *scikit-learn* library and sample experiment using *stream-learn* [11] package.

The presented ensemble is using *Adaptive Windowing* (ADWIN) [12] drift detectors and *Multilayer Perceptron* (MLP) classifiers. ADWIN uses the classifier's output to measure the error rate during the detection process. Ensemble diversity is provided by initializing MLP classifiers with different weights. The

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

corresponding detectors of the ensemble use the individual classifier's predictions. Ensemble integration is handled based on the members' detection from a certain number of previous chunks.

Figure 4 shows the performance of the drift detection ensemble on the generated streams. The black points in the first row present the obtained detections, and the gray points below show the detections of the ensemble's members. The blue dotted lines indicate a single classifier's balanced accuracy score. The red vertical lines mark a moment of drift occurrence.

It can be noticed that the difficulty of streams in terms of drift detection task depends not only on the type of drift but also on the static dataset based on which the stream was generated. For *cubic* drift, the concept change detection seems to state a more difficult task. It is noticeable that some streams are more complex in terms of detection tasks than others.

The occurrence of errors in detection offers the possibility of future improvement – it proves that the generated streams constitute a research challenge and can be used to propose new methods.

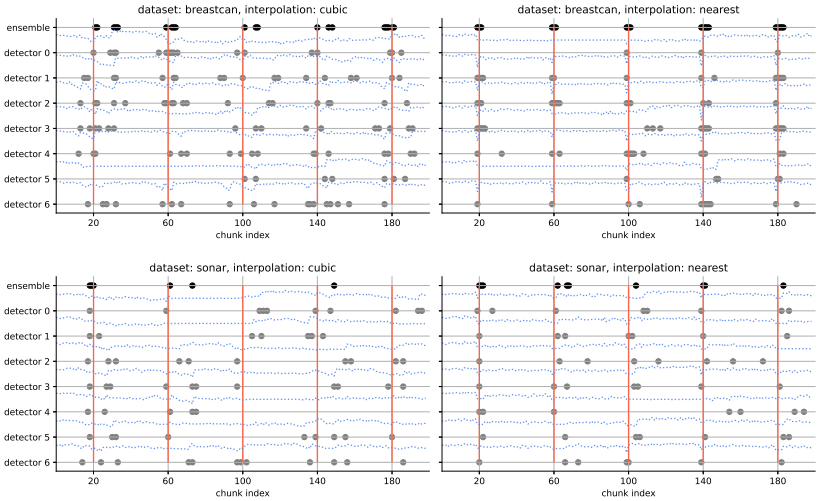

Fig. 4: Drift detection ensemble performance on generated streams of type *cubic* and *nearest*

5 Conclusions

This document proposes a method of data stream generation from real-world static data. The method provides the advantages of synthetic data stream generation – such as the ability to specify stream parameters – while preserving real-world concepts. The parameters of the generated stream can be altered according to the research needs allowing extensive evaluation. In order to gen-

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

erate concept drifts, one-dimensional interpolation of individual data concepts is performed.

The generated streams were used to carry out an exemplary experiment on drift detection ensemble. Analyzing the generated data streams showed their potential in evaluating methods focused on processing data streams. The generator can potentially increase the availability of real-concept stream data with drifts ground-truth.

6 Acknowledgments

This work was supported by the Polish National Science Centre under the grant No. 2017/27/B/ST6/01325 as well by the statutory funds of the Department of Systems and Computer Networks, Faculty of Information and Communication Technology, Wroclaw University of Science and Technology.

References

[1] Maroua Bahri and et al. Data stream analysis: Foundations, major tasks and tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(3):e1405, 2021.

[2] Sergio Ramírez-Gallego and et al. A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, 239:39–57, 2017.

[3] João Gama and et al. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

[4] Jesus L. Lobo and et al. Drift detection over non-stationary data streams using evolving spiking neural networks. In *Intelligent Distributed Computing XII*, pages 82–94, Cham, 2018. Springer International Publishing.

[5] Paweł Zyblewski, Paweł Ksieniewicz, and Michał Woźniak. Classifier selection for highly imbalanced data streams with minority driven ensemble. In *Artificial Intelligence and Soft Computing*, pages 626–635, Cham, 2019. Springer International Publishing.

[6] Albert Bifet and et al. New ensemble methods for evolving data streams. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 139–148, 2009.

[7] Katarzyna Stapor and et al. How to design the fair experimental classifier evaluation. *Applied Soft Computing*, 104:107219, 2021.

[8] Jie Lu and et al. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363, 2018.

[9] Albert Bifet. Classifier concept drift detection and the illusion of progress. In *Artificial Intelligence and Soft Computing*, pages 715–725, Cham, 2017. Springer International Publishing.

[10] Rafael Gonzalez, Richard Woods, and Barry Masters. Digital image processing, third edition. *Journal of biomedical optics*, 14:029901, 03 2009.

[11] Paweł Ksieniewicz and Paweł Zyblewski. stream-learn–open-source python library for difficult data stream batch analysis. *arXiv preprint arXiv:2001.11077*, 2020.

[12] Albert Bifet and Ricard Gavalda. Learning from time-changing data with adaptive windowing. volume 7, 04 2007.