Concept Drift Complexity for Assessing Sampling-Induced Concept Drift in Class-Imbalanced Data Streams

Hajmohammed, M., Chountas, Panagiotis and Chaussalet, Thierry J.

This is a copy of the author's accepted version of a paper subsequently published in the proceedings of the 2024 IEEE 12th International Conference on Intelligent Systems (IS). Varna, Bulgaria 29 - 31 Aug 2024, DOI:10.1109/is61756.2024.10705246.

The final published version will be available online at:

https://doi.org/10.1109/is61756.2024.10705246

Proceedings of 2024 IEEE 12th International Conference on Intelligent Systems (IS)

Concept Drift Complexity for Assessing Sampling-Induced Concept Drift in Class-Imbalanced Data Streams

Mustafa Hajmohammed
School of Computer Science & Eng.
University of Westminster
London, United Kingdom
m.hajmohammed@westminster.ac.uk

Panagiotis Chountas
School of Computer Science & Eng.
University of Westminster
London, United Kingdom
p.i.chountas@westminster.ac.uk

Thierry J. Chaussalet
School of Computer Science & Eng.
University of Westminster
London, United Kingdom
chausst@westminster.ac.uk

***Abstract*—Resampling data is an engineering technique that has an impact on instances where the underlying data distribution and proportions of instances and classes change as a result. Applying any resampling technique may influence the occurrence of certain phenomena such as concept drift, class imbalance, and anomalies. Such influence may produce, exaggerate or eliminate the presence of these phenomena, whether they are viewed as a problem or as a characteristic of the data. Resampling, such as over- or under-sampling, introduces new challenges as well as resolving others. One of the challenges of resampling is its impact on concept drift in a data stream. This paper looks at concept drift produced as a result of data resampling, its nature and how to use its complexity as an indicator of performance. Additionally it examines the nature of concept drifts as a result of applying over- and under-sampling techniques and the various different concept drifts produced as a result of these two techniques. Even though concept drift, class imbalance, and resampling techniques have been studied and researched extensively, sampling-induced concept drift itself as a separate phenomenon has been under-researched. This phenomenon has a certain complexity and can have an impact on the model, which can be measured using concept drift complexity especially when using the value of complexity as a baseline for the overall complexity of drift in a dataset.**

***Index Terms*—concept drift complexity, concept drift, sampling-induced concept drift, sampling techniques, data streams**

I. Introduction

Class imbalance and concept drift are two problems that are not inherently dependent on, yet as this research shows, influence each other. When both are present simultaneously, one phenomenon impacts the treatment of the other [1]. Class imbalance refers to the unequal distribution between the minority and majority classes within a dataset [7], the degree of which varies until equality is achieved. On the other hand, concept drift is defined as the change in the underlying distribution of the target variable [2], the degree of drift varying until uniformity in the distribution is achieved. Class imbalance pertains to the disproportionate distribution of classes within the target variable, while concept drift involves shifts in the statistical properties of the target variable or the relationships between features and the target variable over time. In essence, both issues are concerned with changes in data characteristics, yet they manifest in distinct ways. Although both problems, class imbalance and concept drift, arise due to changes in the distribution of data or relationships within the data, they stem from different causes.

Influencing or manipulating the distribution of a sample or dataset, be it to balance the ratio of classes or any other reason, will collaterally change the distribution of the entire sample or population.

Applying resampling techniques on class-imbalanced datasets, either through over- or under-sampling, influences the distribution of features and target variables providing that learning methods are to be carried out on the resampled dataset.

Online learning, where data is arriving anytime and from anywhere, presents new challenges such as concept drifts, imbalanced data, and anomalies. As opposed to static data, an instance of data often has a single chance to pass through the steps of data preparation and data understanding [9], giving the application less time to react to transient changes, including concept drift and class imbalance.

Concept drift complexity is the degree of severity, longevity and frequency of concept drift in a data stream. In our previous research, we have proposed and demonstrated the benefits of a measuring method for concept drift complexity and its use in baselining and assessing the impact of concept drift overall [6].

Resampling involves two types of sample engineering, over-sampling and under-sampling, each of which may generate one or more different types of concept drift:

- Over-sampling-induced concept drift: artificially generated drift as a result of over-sampling, if measured on the resampled dataset, regardless of the dataset.
- Under-sampling-induced concept drift on the hand consists of new concept drifts due to the elimination of data instances from the dataset changing the statistical properties of the dataset in its entirety.

979-8-3503-5098-2/24/$31.00

In this paper, to complement the work and progress of the research community to address concept drift and class imbalance in data streams, we examine the impact of (re-)sampling-induced concept drift through measuring the complexity and impact of artificially generated concept drift, or lack thereof, its overall impact on the model and ways to assess it. Particular attention will be paid to instances, in which concept drift disappears from the sampled subset or shifts to another subset as a result of a change in data distribution. Prior knowledge of drifting data samples facilitates the development of strategies for addressing resulting issues.

The remainder of the paper is structured as follows: Section II discusses the background, related work, and research pertaining to sampling-induced concept drift. Section III outlines and discusses sampling-induced concept drift. Additionally, Section IV presents a detailed account of the experiments and results conducted in this study. Finally, Section V concludes the paper by offering recommendations for future research and directions in the realm of data distribution arising from sampling rather than actual drifts.

II. Background and Related Work

This paper examines sampling-induced concept drift, the result of applying (re-)sampling techniques to an imbalanced dataset. Sampling-induced concept drift refers to a specific type of concept drift in machine learning where changes in the data distribution arise as a result of data engineering through resampling rather than actual changes in the underlying process generating the data.

It can be argued that the nature of the drift in concept generated from resampling is artificial and/or results from engineering of the data as opposed to concept drifts resulting organically during data gathering.

[5] argues that the use of concept drift detectors is virtually non-existent in imbalanced data streams. However, the reader may find many examples in machine learning where concept drift detectors have been used in data streams. In such a context, it is acceptable to some extent that most detectors use adaptability and agility to handle data streams in incremental learning settings. [7] suggests that datasets should be partitioned into smaller subsets that are ultimately used to form disjoint rules pertaining to class concepts, referring to the break down of datasets into smaller, more manageable subsets.

A sampling shift (drift) is alternatively referred to as a "virtual" drift while a concept shift is defined as a "real" shift [8]. [2] argues that, from a practical standpoint, it is the ability to detect and adapt to concept drift, which is crucial, regardless of its type. The specific nature of the drift (real or virtual) is secondary to the necessity of maintaining the model's accuracy and performance.

Online learning methods pose greater challenges due to the availability of only one instance at each time step [12]. In response to this challenge, the concept drift complexity method has been employed in this research to monitor and evaluate the evolving impact of sampling-induced concept drift, utilizing one instance at each time step throughout the data stream's lifespan.

Adaptive synthetic sampling approaches for imbalanced learning, such as ADASYN [13] and Borderline-SMOTE [14], are designed to generate synthetic data samples near the decision boundaries of two or more classes. These methods can be particularly effective in evolving datasets of data streams, as they adapt the generation of synthetic samples based on newly observed attributes and features of the minority classes. The connection between generating synthetic samples near decision region borders and concept drift lies in the self-adaptability of the model to changes in the characteristics and distribution of samples in data streams, which is crucial for online machine learning.

Detecting and addressing changes in the underlying hidden contexts of data using a single algorithm has proven to be challenging [2], [11] expands upon the argument, highlighting the ongoing research gaps in online imbalance learning, with a particular focus on addressing concept drift.

To the best of our knowledge, no other published research on sampling-induced concept drift and its impact as a topic, or the use of concept drift complexity as an indicator for detection in incremental learning in general and online learning in particular, exists today.

III. Sampling-Induced Concept Drift

The two problems of class imbalance and concept drift are distinctly different in their nature, and the present of either does not necessarily cause the other. However, both phenomena can and do occur simultaneously. As will be demonstrated in the experiments in this paper, the impact of applying class imbalance sampling, either over-sampling or under-sampling, can impact the nature of a running drift in many ways from exaggerating it, shifting it to a successive sample (sampling shift), masking it, to eliminating it altogether.

Concept drift can be influenced when instances of the target variable are altered by resampling the class labels, either increasing or decreasing their occurrences. In such cases, there are three potential outcomes: a) the target variable may become excessively replicated generating or shifting concept drifts, b) it may be under-replicated leading to fewer concept drifts occurring, or c) the sampling process could eliminate drift by uniformly reducing differences within the distribution.

In the view of Wang et al. "online [...] learning often combines the challenges of both class imbalance and concept drift" [1], a view that fundamentally supports the argument that concept drift is indeed an online-learning problem in imbalanced data(sets). Such a view is, in principle, in agreement with this research's question, and yet an inter-dependency between the two problems does not necessarily exist.

The concept drifts that appear as a result of inappropriate sampling are of a temporal nature. In a data stream, applying sampling techniques on a batch of data or a subset of a dataset is by definition a transient process; its characteristics and distributions are local to the instances on which the sampling is applied, therefore it is safe to assume that applying

sampling may or may not be applicable to the next batch of instances. For example, applying random resampling on historical temperature data in a weather forecasting model can introduce abrupt or gradual artificial changes in temperature predictions, causing the model to incorrectly forecast sudden and extreme weather fluctuations (concept drift).

When over-sampling is applied, especially in a significant manner, the class distribution in the training dataset no longer reflects the true distribution of the classes in the real-world data. For example, if a minority class is over-sampled to match the number of instances of the majority class, the model learns from a dataset where rare events are as frequent as common ones, which leads to virtual drift. Virtual concept drift is a sampling shift that does not have an impact on the decision boundary [2], with no immediate impact on learning performance, relative to the model.

IV. Experiments

Three experiments were conducted to explore and examine sampling-induced concept drift in an imbalanced data stream:

(A) A self-generated dataset following a normal (Gaussian) distribution, providing evidence that class imbalance sampling can, indeed, create concept drift.

(B) A self-generated dataset following a normal (Gaussian) distribution, showing the nature of concept drift as a result of over- and under-sampling on an imbalanced data stream.

(C) The evaluation and measuring of concept drift complexity and the overall cumulative impact of concept drift in a data stream.

The experimental reasoning behind using a self-generated dataset is that the nature of the dataset does not affect the data distribution, class labels, or the nature of class imbalance. Both self-generated and real-world datasets would exhibit similar characteristics and levels of imbalance, regardless of their source. Hence, we decided to create the data for the experiments synthetically instead of using a real-world example, as the experiments and results would remain comparable in either case.

A. Concept Drift Through Sampling using SMOTE

Experiments were conducted using a population of 30 000 instances to predict loan defaults whose class imbalance ratio is 7 500 for the minority class and 22 500 for the majority class as shown in Table I, i.e. 25% of loan holders default. The binary value of the target class ”Default” signifies whether a loan default has occurred. For the purpose of this experiment, concept drift has been purposefully designed to happen in one feature, employment length, where shorter duration of employment can predict defaults on loans using the following concept drift and class imbalance analysis:

- Concept drift: changing employment situations (i.e. average employment duration) directly affect the feature distribution and could influence default rate predictions.
- Class imbalance: with a shift towards shorter employment duration, default rates might rise. Class imbalance may arise if default rates increase disproportionately compared to non-default rates.

TABLE I: Concept Drift Analysis

	Amount	Interest (%)	Emp. Length in Years	Default
0	29967	3.68	10.5	0
1	23617	5.77	9.9	0
2	31477	5	1.8	1
3	40230	3.46	11.5	0
4	22658	5.49	9.8	0
5	47438	4.13	2.1	1

In this experiment, to balance the class distribution of the dataset, we employed SMOTE (Synthetic Minority Over-sampling Technique) [3], which increases the number of minority instances by generating synthetic samples rather than duplicating them. The results indicate that after applying over-sampling to the dataset, concept drift shifts.

To evaluate the effect of sampling on a class-imbalanced dataset, we applied SMOTE for over-sampling and compared the results before and after. To identify any changes in rate of concept drift within the dataset, we utilised the ADWIN (ADaptive WINdowing) [4] algorithm for concept drift detection.

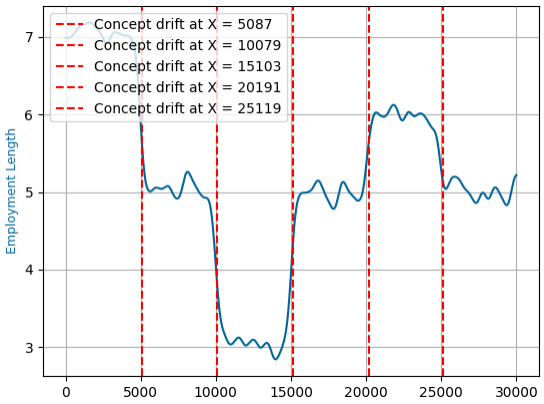

Fig. 1: Concept Drift Before Applying Sampling

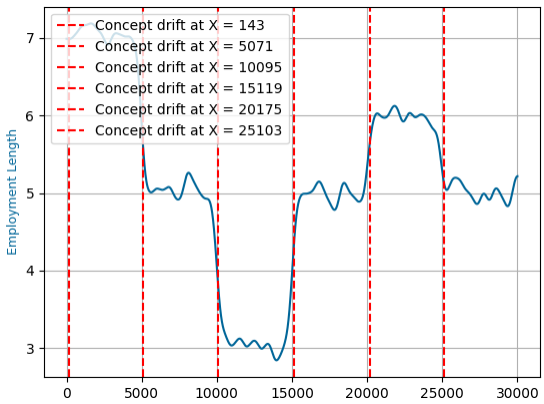

Fig. 2: Concept Drift After Applying Sampling

In both experiments, concept drift presented an issue both before and after resampling, with the following observations: In the first experiment, before applying any sampling to the data, concept drift appeared five times as illustrated in Fig. 1. After over-sampling the dataset, a new concept drift appeared at point X = 143 as illustrated in Fig. 2, whereas the value of X represents the number of instances, indicating that the over-sampling caused the underlying data distribution to change, introducing a new concept drift.

In Fig. 2, the results indicate the emergence of one real concept drift due to over-sampling, along with five virtual concept drifts. Interestingly, while the original dataset displayed five concept drifts prior to resampling, these still appeared to undergo shifts.

For these existing concept drift(s), the underlying data distribution of the whole dataset has shifted, indicating that there is an impact on the entire population of the stream as a result of over-sampling. Any shift, however small, indicates that the drift may shift to the next subset of data, impacting it more significantly.

B. Concept Drift Through RandomSampling

In the second set of experiments, we applied RandomOver-Sampling and RandomUndersampling [10] to the same dataset used in the first experiment. While the focus of the second experiment is on under-sampling, we evaluated the results of RandomOverSampling against those of the first experiment.
Before RandomUnderSampling:

- The initial class distribution was as follows: Class 0 (representing no loan default) had 22 500 samples, while Class 1 (representing loan default) had 7 500 samples, indicating a significant class imbalance.

After RandomUnderSampling:

- The adjusted class distribution post-undersampling shows an equalised representation of both classes, with 7 500 instances each for class 0 and class 1 (representing default and non-default).

TABLE II: Concept Drift Analysis

Stage	Concept Drift	Emp. Length
Before RandomUnderSampling	5087	5.5
	10079	2.0
	15103	3.5
	20191	6.0
	25119	5.1
After RandomUnderSampling	4335	5.1
	7631	6.3
	8847	4.4
	10095	3.9
	11311	7.7
	12687	5.8
	13839	5.6

As the results in Table II indicate, the impact of resampling has not only shifted concepts but created new instances of concept drift. These analyses indicate the presence of concept drift in the post-resampling dataset, highlighting the importance of ongoing monitoring and adaptation of the model to account for changes in the data distribution over time. These analyses do not indicate interdependency between class imbalance and concept drift, but rather that the impact of handling class imbalance on concept drift is significant.

RandomOverSampling on the other hand did not result in as significant a change in concept drifts. As Table III shows, the initial concept drifts in the dataset, despite shifting slightly, remained present in the analysed post-resampling dataset, while new drifts appeared earlier than present in the dataset pre-resampling such as the concept drift at X = 143, agreeing with the result of the first experiment.
Before RandomOverSampling:

- The classes were highly imbalanced, with a ratio of 75:25 between the majority (class 0 no loan default) and minority (class 1 default) classes.

After RandomUnderSampling:

- The classes are balanced, with both classes having equal representation.

TABLE III: Concept Drift through RandomOverSampling

Stage	Index	C. Drift	Emp. Length
Before RandomOverSampling	5087	Detected	5.5
	10079	Detected	2.0
	15103	Detected	3.5
	20191	Detected	6.0
	25119	Detected	5.1
After RandomOverSampling	143	Detected	7.2
	5071	Detected	4.8
	10095	Detected	4.6
	15119	Detected	7.1
	20175	Detected	5.8
	25103	Detected	6.1
	34319	Detected	7.0

Impact of resampling on concept drift:

- Both over-sampling and under-sampling affect where and when concept drift is detected.
- The altered detection points indicate changes in data patterns due to different resampling techniques.
- Each method introduces different biases and sensitivities to the dataset, affecting model performance and drift detection.

C. Measuring Concept Drift Complexity

In this experiment, we applied over-sampling to a dataset with 30 000 instances to address class imbalance, where the ratio was 75% true and 25% false, corresponding to 22 500 true instances and 7 500 false instances. After resampling,

the dataset comprised 45 000 samples, resulting in balanced classes. We measured the concept drift complexity ($comp_N = \max\{comp_i, comp_{i-1}, ..., comp_0\} + comp_i$) [6] on the entire dataset to examine the impact of concept drift on predictive performance. As shown in the second experiment in Table III, there was a minor shift in concept drift as a result of the random over-sampling. However, this impact does not appear to affect the dataset significantly, as shown in Fig. 3. Concept drift complexity has been used as a metric for measuring performance degradation due to resampling techniques.

Fig. 3: Overall Complexity and Impact of Concept Drift on Performance

The value of concept drift complexity showed a slight increase as the data samples arrived, indicating that the impact of concept drifts on the dataset is negligible, despite their presence after resampling. Furthermore, the value of concept drift complexity appears to remain close to the Mean Complexity Score $\bar{x}$ as more data instances arrive. This interpretation of the results suggests that the impact on the performance of the predictive models on the dataset remains intact, and no degradation in learning occurred. Had the concept drift complexity end value increased excessively or deviated significantly from the mean concept drift $\bar{x}$, it would have indicated that the impact of concept drifts in the resampled dataset had a much bigger effect on the model and its performance. These results provide the observer with a means to use concept drift complexity to evaluate a model's performance in the presence of concept drift in data streams.

V. Conclusion and Future Works

In this paper, we examined the impact of two different techniques for treating class imbalance on concept drift and the nature of sampling-induced concept drift.

We have been able to demonstrate that resampling, specifically over- and under-sampling are capable of introducing not only new concept drifts not present in the dataset pre-resampling but also shifts in pre-existing drifts. Furthermore, our research shows that measuring complex drift complexity can show both the impact of concept drift on a dataset and the temporal shifts that occur as a result of resampling.

Additional research will be required to identify effective mitigation measures for the phenomena described in this paper and/or into improved sampling and resampling techniques for data streams that should, ideally, avoid the issue of sampling-induced concept drift altogether or, at the very least, minimise it.

References

[1] S. Wang, L. L. Minku and X. Yao, "A Systematic Study of Online Class Imbalance Learning With Concept Drift," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 10, pp. 4802-4821, Oct. 2018, doi: 10.1109/TNNLS.2017.2771290.

[2] A. Tsymba, "The Problem of Concept Drift: Definitions and Related Work", Department of Computer Science, Trinity College, 2004, Available: https://www.scss.tcd.ie/publications/tech-reports/reports.04/TCD-CS-2004-15.pdf

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique", Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.

[4] A. Bifet and R. Gavalda`, "Learning from Time-Changing Data with Adaptive Windowing," Proceedings of the 2007 SIAM International Conference on Data Mining, Apr. 2007, doi: https://doi.org/10.1137/1.9781611972771.42.

[5] Ł. Korycki and B. Krawczyk, "Concept Drift Detection from Multi-Class Imbalanced Data Streams," 2021 IEEE 37th International Conference on Data Engineering (ICDE), Chania, Greece, 2021, pp. 1068-1079, doi: 10.1109/ICDE51399.2021.00097.

[6] M. Hajmohammed, P. Chountas, T. J. Chaussalet, "A Concept Drift Based Approach To Evaluating Model Performance And Theoretical Lifespan," unpublished.

[7] H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

[8] M. Salganicoff, "Tolerating Concept and Sampling Shift in Lazy Learning Using Prediction Error Context Switching," Springer eBooks, pp. 133–155, Jan. 1997, doi: 10.1007/978-94-017-2053-3-5.

[9] R. Elwell and R. Polikar, "Incremental Learning of Concept Drift in Nonstationary Environments," in IEEE Transactions on Neural Networks, vol. 22, no. 10, pp. 1517-1531, Oct. 2011, doi: 10.1109/TNN.2011.2160459.

[10] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Online learning from imbalanced data streams," in Proc. Int. Conf. Soft Comput. Pattern Recognit., 2011, pp. 347–352

[11] K. Malialis, C. G. Panayiotou and M. M. Polycarpou, "Online Learning With Adaptive Rebalancing in Nonstationary Environments," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 10, pp. 4445-4459, Oct. 2021, doi: 10.1109/TNNLS.2020.3017863.

[12] H. Zhang, W. Liu and Q. Liu, "Reinforcement Online Active Learning Ensemble for Drifting Imbalanced Data Streams," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 8, pp. 3971-3983, 1 Aug. 2022, doi: 10.1109/TKDE.2020.3026196.

[13] Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

[14] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," Lecture Notes in Computer Science, pp. 878–887, 2005. doi:10.1007/11538059-91