

A Critical Review on Concept Drift Monitoring Process for Class Imbalance in Data Streams

Nouhaila Aasoum^a, Ismail Jellouli and Souad Amjad

Computer Science and Systems Engineering Laboratory, Abdelmalek Essaadi University, Tetouan, Morocco

Keywords: Machine learning, online learning, concept drift, class imbalance, non-stationary environment.

Abstract: Machine learning techniques have participated in world evolution. They have accomplished worthy goals in many areas such as banking, industry, cybersécurité, and many others. However, in most data analysis applications, data comes in streams based on online learning scenarios. As streams emerge and change quickly over time, it will be hard to store them in memory. Thus, the analysis has become a real challenge to mitigate using traditional approaches. The change in data distribution degrades the accuracy performance of the trained model and becomes inefficient. This phenomenon is called concept drift, where the model must adapt quickly to these changes, including those in the environment, trends, or behaviour, to maintain their accuracy. Another phenomenon commonly exists in real-world applications is a class imbalance, when data distribution changes within classes. Thus, the model will favor the majority class and ignores the minority one. The problem becomes more challenging when both of them co-exist. Therefore, a few studies addressed this research gap. The objective is to detect concept drift with class imbalance for enhanced performance in a non-stationary data environment. This study will focus on class imbalance handling techniques, and concept drift effects in decreasing model performance, and some methods to detect concept drift while existing class imbalance issue.

1 INTRODUCTION

Learning from streaming data is an emerging topic that has received much attention in the last decades. In many machine learning applications, data is generated instantly over time without storing it as static data. Thus, we have to use a real-time dataset to build a suitable online classifier that maintains its performance in non-stationary environments.

The non-stationary environment is characterized by newly emerged samples of data, representing the minority portion of the entire dataset. The minority class contains a few instances, while those instances can be essential for the trained model.

However, this model focuses on the majority class and ignores the minority one. Training from such imbalanced streaming data is called online class imbalance (S. Wang et al., 2013).

For example, figure 1 illustrates class imbalance for spam email filtering, where the non-spam email presents the majority class, while the spam email shows the minority class from the entire dataset.


Another challenging problem is concept drift that degrades the model accuracy and makes it unsuitable because of the changes in the data distribution.

The model needs to adapt quickly to such changes if any features or classes emerge in the data. The concept drift becomes a severe issue when the changes affect sample features and class numbers simultaneously. Thus, it becomes hard to handle it.

When class imbalance and concept drift coincide, the problem will be more challenging because one issue can affect the treatment of the other, especially in online scenarios.

2 CLASS IMBALANCE MANAGED TECHNIQUES

The class imbalance issue has received much attention recently. Many state-of-the-art propose techniques in this research gap for both stationary and non-stationary environments to handle it by focusing

^a <https://orcid.org/0000-0002-7283-6785>

on data preprocessing, algorithmic, and ensemble learning approaches.

2.1 Data Preprocessing Approaches

Resampling is a preprocessing technique that is used before training the model in real-time.

The objective is to maintain the number of instances for each category to benefit from the ignored category distributions. It works autonomously with the learning algorithm at the data level (S. Wang et al., 2015). The principal used types of resampling are oversampling and undersampling. Oversampling-based Online Bagging (OOB) (S. Wang et al., 2015) is a technique that increases the minority class. Undersampling-based Online Bagging (UOB) (S. Wang et al., 2015) is a technique that decreases the majority class. While relatively, each has significant shortcomings: (UOB) ignores data from the majority class, whether this data is necessary or not. (OOB) generates exact samples from the minority class, which may produce overfitting in the classifier (Ditzler & Polikar, 2010).

(Chawla et al., 2002) have proposed synthetic Minority Oversampling Technique (SMOT). It is an oversampling method that generates new samples from the minority class, depending on the features of the nearest neighbours. Thus, SMOT maintains the accuracy of the minority class in classification tasks compared to other approaches. There are different used techniques like random oversampling and random undersampling.

2.2 Algorithmic and Ensemble Approaches

Ensemble approaches manage the issue of imbalanced class differently. They focus on increasing the accuracy of the minority category by updating some training mechanisms.

(J. Wang et al., 2012), and (Ghazikhani et al., 2013) proposed online cost-sensitive learning methods CSOGD, RLSACP, respectively. They misclassified the cost of the algorithm and made him capable of favouring the minority class.

Additionally, there is an ensemble learning approach based on multiple classifiers. It combines the classification of various classifiers and trains them individually. Bagging or bootstrap aggregating (Oza & Russell, 2001) uses an ensemble-based approach by combining multiple classifiers to improve the model accuracy. In bagging, we can mitigate the imbalance by generating more test data. Boosting is another approach based on ensemble classifier, the

procedure in this method is different. In boosting, we give much importance to the minority class. Where taking a subset from the entire dataset misclassifies the samples.

Generally, bagging reaches to reduce the model variance while boosting reaches for improving the model accuracy.

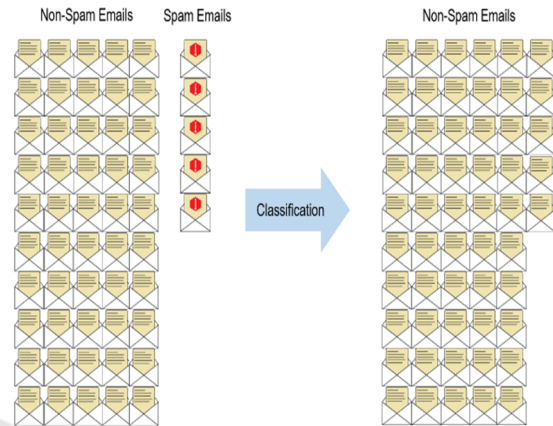


Figure 1: Imbalanced data in spam email filtering.

3 CONCEPT DRIFT TAXONOMY

Data comes from the same source distribution in stationary environments, but in real-world applications, data comes from different sources and arises concept drift phenomenon when existing online learning scenarios. This issue can directly affect the accuracy of the trained model and decrease their performance in classification or prediction results.

In principle, for supervised tasks, the model needs inputs and outputs features from the data. However, these features can change, or classes number can increase over time. Thus, we have to separate between virtual and real concept drift. The equation below shows the Bayesian decision theory (Jameel et al., 2020).

$$P(c/X) = P(c) P(X/c) / P(X) \tag{1}$$

Where posterior, prior, conditional, and feature-based probabilities are $P(c/X)$, $P(c)$, $P(X/c)$, and $P(X)$, respectively (Jameel et al., 2020). Real drift mentions the $P(c/X)$ changes that affect the probability of a class label and given features. This change can decrease the accuracy of the classifier and affects decision boundaries. (Wares et al., 2019).

Alternatively, virtual drift refers to $P(X)$ changes only. In this state, the distribution has changed without affecting the classifier's decision limits. (Wares et al., 2019). For the Hybrid drift, the changes

affect $P(X)$ and $P(c/X)$ simultaneously. Figure 2 shows the differences between drift types:

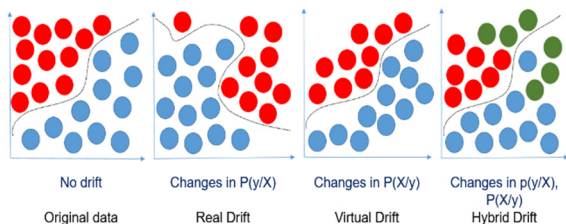


Figure 2: Concept drift types.

However, concept drift has a different configuration pattern occurrence within the data streams over time.

Abrupt: When the change occurs suddenly, in this case, another target can replace the main one.

Incremental: The changes are relatively, and they may take a lot of time to compare.

Gradual: In this type, the change is replaced progressively from an interest to another one and can be tracked over a long time. The gradual drift is the most challenging type to detect. For example, when a sensor is becoming old and not providing correct outputs in a suitable time. (Gözüaçık & Can, 2020).

Re-Occurring: This type of drift is concerned with seasonal changes that can revert after a few times. Moreover, this matter occurs when there is demand for a particular thing or event during the year or month. Figure 3 illustrates all configuration patterns of concept drift.

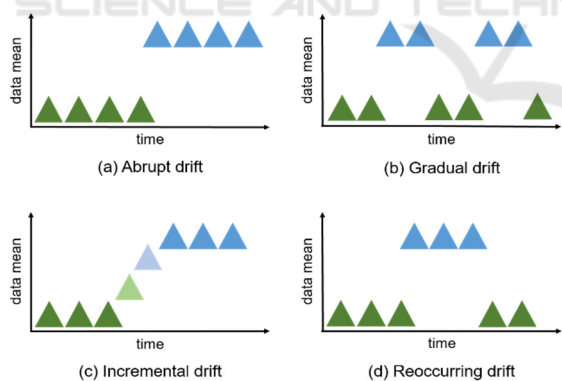


Figure 3: Concept drift configuration patterns.

4 APPROACHES TO DETECT CONCEPT DRIFT WITH CLASS IMBALANCE

A few detection methods have been submitted to detect concept drift for class imbalance in online scenarios through the existing studies.

(S. Wang et al., 2013) proposed Drift Detection Method for Online Class Imbalance DDM-OCI is a helpful technique for handling drift in the minority class, based on tracking his recall. At the same time, it is a more robust and suitable technique for imbalanced data streams.

(H. Wang, 2015) proposed another method is the Linear Four Rate (LFR). It is an improvement approach for DDM-OCI. LFR is capable of detecting drift with the best recall and few false alarms. Its controls the four rates; precision and recall for both the minority and the majority class.

This study (Mirza et al., 2015) proposed Ensemble of Subset Online Sequential Extreme Learning Machine (ESOS-ELM). It is an ensemble-based technique that works for non-stationary environments. In ESOS-ELM, a single classifier processed a majority class instance, while multiple classifiers process the minority class. In this case, classifiers learned from balanced subsets of the original dataset that is imbalanced.

Page-Hinkley (PH) (Brzezinski & Stefanowski, 2015) proposed drift detecting, especially for the abrupt one, addressed for the binary imbalance issue only. A recent drift detector submitted from multiclass imbalanced data streams in this study (Krawczyk, 2021), based on Restricted Boltzmann Machine (RBM-IM) notify if any data is drifting. It is capable of training itself dynamically and detects the drift at a local and global level. RBM-IM can manage the changes in minority class without being biased towards the majority one.

CALMID is a comprehensive active learning method using for multiclass imbalanced where existing concept drift, proposed by (Liu et al., 2021). It's a recent approach based on an ensemble classifier. CALMID considers the multiclass imbalance ratio as unbiased. When the sensor reports any drift, a new classifier will be generated and initialized by the existing weighted at the initialization training (Liu et al., 2021). Thus, it is a suitable approach when concept drift and multiclass imbalance co-exist.

5 COMPARATIVE STUDY

This section presents a comparative study of different detector approaches for imbalanced data streams in various studies. The comparative analysis is based on the effectiveness of detecting several types of drifts with the imbalanced class issue.

Table 1: Comparison of concept drift detection approaches for imbalanced data streams.

Method	Class imbalance	Multiclass	Drift type	Ref
DDM-OCI	X	X	Real drift	(S. Wang et al., 2013)
LFR	X	X	Real drift	(H. Wang, 2015)
PH	X	X	Real drift	(Brzezinski & Stefanowski, 2015)
ESOS-ELM	✓	X	Real drift, Virtual drift	(Mirza et al., 2015)
RBM-IM	✓	✓	Real, Virtual and Hybrid drift	(Krawczyk, 2021)
CALMID	✓	✓	Real, Virtual and Hybrid drift	(Liu et al., 2021)

6 CONCLUSION

This paper has discussed concept drift with class imbalance for online learning, focusing on class imbalance mitigating techniques. We can state that class imbalance handling techniques are still not applicable for concept drift detection through the existing studies. Then we have talked about the most-used methods for concept drift detection for imbalanced data streams.

According to the literature, a few studies have been proposed when both issues co-exist. In addition, this is due to the difficulty that they arise in online scenarios. The majority of proposed methods do not cover all concept drift types (virtual drift, real drift, and hybrid drift).

Thus, there is no one method for all in this research gap. We can conclude that concept drift detection approaches need to be more adaptive and applicable with their different types, mainly when it comes to online scenarios where data change by its nature.

REFERENCES

- Brzezinski, D., & Stefanowski, J. (2015). Prequential AUC for classifier evaluation and drift detection in evolving data streams. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 8983, 87–101. https://doi.org/10.1007/978-3-319-17876-9_6
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(February 2017), 321–357. <https://doi.org/10.1613/jair.953>
- Ditzler, G., & Polikar, R. (2010). An ensemble based incremental learning framework for concept drift and class imbalance. *Proceedings of the International Joint Conference on Neural Networks, August*. <https://doi.org/10.1109/IJCNN.2010.5596764>
- Ghazikhani, A., Monsefi, R., & Sadoghi Yazdi, H. (2013). Recursive least square perceptron model for non-stationary and imbalanced data stream classification. *Evolving Systems*, 4(2), 119–131. <https://doi.org/10.1007/s12530-013-9076-7>
- Gözüaçık, Ö., & Can, F. (2020). Concept learning using one-class classifiers for implicit drift detection in evolving data streams. *Artificial Intelligence Review*, 0123456789. <https://doi.org/10.1007/s10462-020-09939-x>
- Jameel, S. M., Hashmani, M. A., Alhussain, H., Rehman, M., & Budiman, A. (2020). A critical review on adverse effects of concept drift over machine learning classification models. *International Journal of Advanced Computer Science and Applications*, 11(1), 206–211. <https://doi.org/10.14569/ijacsa.2020.0110127>
- Krawczyk, B. (2021). *Concept Drift Detection from Multi-Class Imbalanced Data Streams*. April.
- Liu, W., Zhang, H., Ding, Z., Liu, Q., & Zhu, C. (2021). A comprehensive active learning method for multiclass imbalanced data streams with concept drift. *Knowledge-Based Systems*, 215, 106778. <https://doi.org/10.1016/j.knosys.2021.106778>
- Mirza, B., Lin, Z., & Liu, N. (2015). Ensemble of subset online sequential extreme learning machine for class imbalance and concept drift. *Neurocomputing*, 149(Part A), 316–329. <https://doi.org/10.1016/j.neucom.2014.03.075>
- Oza, N. C., & Russell, S. (2001). Experimental comparisons of online and batch versions of bagging and boosting. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 359–364.

- <https://doi.org/10.1145/502512.502565>
- Wang, H. (2015). *Concept Drift Detection for Streaming Data t The three user defined parameters are the time decaying*.
- Wang, J., Zhao, P., & Hoi, S. C. H. (2012). Cost-sensitive online classification. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 1140–1145. <https://doi.org/10.1109/ICDM.2012.116>
- Wang, S., Minku, L. L., Ghezzi, D., Caltabiano, D., Tino, P., & Yao, X. (2013). Concept drift detection for online class imbalance learning. *Proceedings of the International Joint Conference on Neural Networks*. <https://doi.org/10.1109/IJCNN.2013.6706768>
- Wang, S., Minku, L. L., & Yao, X. (2015). Resampling-based ensemble methods for online class imbalance learning. *IEEE Transactions on Knowledge and Data Engineering*, 27(5), 1356–1368. <https://doi.org/10.1109/TKDE.2014.2345380>
- Wares, S., Isaacs, J., & Elyan, E. (2019). Data stream mining: methods and challenges for handling concept drift. *SN Applied Sciences*, 1(11), 1–19. <https://doi.org/10.1007/s42452-019-1433-0>

