Proceedings of Machine Learning Research 183:187–198, 2022 Short Paper - LIDTA 2022

The Influence of Multiple Classes on Learning from Imbalanced Data Streams

Agnieszka Lipska AGNIESZKA.L1995@GMAIL.COM
Jerzy Stefanowski JERZY.STEFANOWSKI@CS.PUT.POZNAN.PL
Institute of Computing Science, Poznan University of Technology, 60-965 Poznań, Poland

Editor: Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak and Shuo Wang.

Abstract

This work is aimed at examining the influence of local data characteristics and drifts on the difficulties of learning online classifiers from multi-class imbalanced data streams. The results of many experiments with synthetically generated data streams have shown a much greater role of the overlapping between many minority classes (the type of borderline examples) than for streams with one minority class. The presence of rare examples in the stream is the most difficult single factor. Unlike binary streams, the specialized UOB and OOB classifiers perform well enough for even high imbalance ratios. The most challenging for all classifiers are complex scenarios integrating many drifts and factors simultaneously, which worsen the evaluation measures stronger than for binary ones.

Keywords: Multiple classes in imbalanced data, data streams, local difficulty factors

1. Introduction

Learning classifiers in the context of non-stationary data streams has been intensively studied in the last twenty years. Besides new computational requirements, it is still challenging to deal with *concept drifts*. This task becomes even more challenging in the presence of additional data complexities, in particular imbalances between classes. However, the current research on imbalanced and concept drifting streams are not as developed as in the case of separately considered static data or streams. The number of new algorithms for dealing with imbalanced stream classification is still much smaller. Moreover, existing works mostly focus on re-balancing classes and reacting to changes affecting the global imbalance ratio. These works do not sufficiently consider more complex imbalanced streams scenarios, where these changes are accompanied by the *local difficulty factors* already considered for the static imbalanced data such as: the minority class fragmentation into sub-concepts, class overlapping or occurrence of different types of unsafe minority examples (borderline, rare or outliers (Napierała and Stefanowski, 2016)). In evolving data streams these factors could also influence the changes in *local class distributions* and other *local drifts*. Nevertheless, the conjunctions of these data factors and drifts have not been sufficiently studied yet, see discussions in (Aguiar et al., 2022; Brzezinski et al., 2021).

A new categorization of concept drifts with data difficulty factors for class imbalanced streams is introduced in (Brzezinski et al., 2021). These authors also carried out comprehensive experiments with synthetic and real data streams showing the different influence of types of minority examples and class decomposition on predictions of representative online

classifiers. Combinations of multiple factors demonstrated to be the most challenging for classifiers, which were not able to recover from these drifts.

Nevertheless (Brzezinski et al., 2021) is limited to considering binary imbalanced classes only. In the current paper we carry out new experiments with studying the aforementioned data factors and drifts for multiple classes streams, which has not been investigated yet.

We want also verify observations from the studying binary streams will be valid also for the multiple classes. Furthermore following other experiments with static multiple data (Lango and Stefanowski, 2022), we are interested to check whether the role of borderline examples will be more important in case of many overlapping classes. Therefore, we decided to create an additional generator of synthetic streams. Unlike experimental studies aimed at comparisons of several classifiers, we claim that this kind of study could support better understanding the difficulties of this type of drifting data streams and at the same time could indicate the directions of development of new specialized algorithms.

2. Categorization of data difficulty factors and local concept drifts

Due to page limits the reader is referred to such books as (Gama, 2010) for basic concepts of data streams and dealing with concept drifts and to (Fernández et al., 2018) for the aspects of static imbalanced data. The imbalance between classes is characterized by the *global imbalance ratio*. Here for multiple classes, these ratios are defined as the percentage of examples in the data that belong to the minority class. In case of static imbalanced data it has been shown that beside that other sources of classifiers deterioration include: (1) the decomposition of the minority class into several sub-concepts, (2) the presence of small, isolated groups of minority examples located deeply inside the majority class region (it corresponds to *rare cases*), (3) the effect of strong overlapping between the classes.

The last two factors can be identified through the so-called *types of examples* (Napierała and Stefanowski, 2016), which distinguished between safe and unsafe examples. *Safe examples* are the ones located in homogeneous regions populated by examples from one class only. The *unsafe* examples are categorized into *borderline* (placed close to the decision boundary between classes), *rare cases* (isolated groups of few examples located deeper inside the opposite class), and *outliers*. Following the method from (Napierała and Stefanowski, 2016) these types of examples can be identified based on the analysis of class labels of other examples in the local neighborhood.

Brzezinski et al. (2021) introduced an *extended concept drift categorization from imbalanced streams* that takes into account these local data factors into account, which covers four main criteria and specific types of drifts inside them. Its first criterion distinguishes the *locality in the data streams*, i.e. the interest is focused on the local bounded sub-regions in the attribute space and changes occurring in these local regions as opposed to global drifts. The second criterion corresponds to a class composition. A *homogeneous* class composition means that examples of the minority class are concentrated in a single local region while a *heterogeneous* class composition means that examples of the class are spread over multiple local regions (which are larger than groups of rare minority examples). It corresponds to mainly drifts being either class region split or merging them. Then, two situations are considered: *static imbalance ratio* if the class proportions do not change over time, and *dynamic imbalance ratio* if the imbalance ratio changes over time. The last criterion is referred to

monitoring proportions of the types of the minority class examples in the stream, which could be changing or static over streams.

3. An experimental setup

The aim of experiments is to investigate the influence of the discussed data difficulty factors and drifts on predictive performance of selected online classifiers. As these experiments refer to the earlier study with binary imbalanced streams, we follow and extend its experimental setup (Brzezinski et al., 2021). In general we want to answer two research questions: [**RQ1**] What is the impact of different types of single data difficulty factors and isolated, single drifts on the classification performance? [**RQ2**] Which complex scenarios integrating several data factors and drifts are the most harmful for classification performance? In this context we want also compare new results for multiple imbalanced classes with earlier ones obtained by binary (minority vs. majority) classes (Brzezinski et al., 2021). Our experiments cover the following difficulty factors and drifts:

- *Imbalanced ratios* – into two scenarios of either static data or with a single drift with the following configurations: one majority class and two minority classes with ratios (sizes referred to all examples) (1% 1%); (3% 3%); (5% 5%); (10% 10%) and the balanced scenario (30% 30%); and with three minority classes (1% 1% 1%); (3% 3% 3%); (5% 5% 5%); (10% 10% 10%) and the balanced version with 25%.

- *Types of minority examples* being either *borderline* (class overlap) or *rare* ones (other examples in the minority classes are generated to be safe ones), i.e. with the given percentages of their occurrence in each class: 20%, 40%, 60%, 80% and 100%.

- *Changes in class composition*, i.e. local drifts of the split of the each minority class into 3, 5 and 7 sub-clusters; and moving these numbers of sub-clusters.

We carried out the experiments in a controlled framework based on synthetic generated data, where each data factor can be modeled and parametrized according to different planned scenarios. We prepared two generators[1]:

1. The *old generator* - the same as used in the previous experiments (Brzezinski et al., 2021), where minority classes are generated in elliptical spheres and the majority class instances uniformly surround them[2].

2. The *new generator* – the majority class has an elliptical shape and its overlaps with the remaining class of also the similar shape.

In further subsections we will abbreviate the old generator by **O** and new one by **N**; in the names of data streams the single factors will denoted by number referring to their values, e.g. imb_0.01_0.01 is a balanced data stream with the static imbalance 1% for two minority class and one majority class.

1. The more detailed descriptions of both generators and their parametrization are given in an electronic appendix https://www.cs.put.poznan.pl/jstefanowski/pub/append-streams.pdf
2. See https://github.com/dabrze/imbalanced-stream-generator for its description and code.

All stationary streams consist of 200,000 generated examples, while drifting streams have 250,000 examples. They are modeled with 5 attributes. The single incremental drift is always started at 70 000 example and ends before 100 000 example. In next sections, we will show results for the following points: after the start of stream, the pre-drift (at 70,000 example), post-drift (100,000 where the lower decrease is usually observed in plots) and at the end of the stream. Unless stated otherwise, for drifting scenarios the minority classes are generated as single clusters of safe-type instances and all classes are balanced.

Classifiers: We chose four different online classifiers - the same as studied in (Brzezinski et al., 2021)[3]. Two of them are not specialised for dealing with imbalanced data streams: *Online Bagging* (OB) ensemble and VFDT single classification tree. The next ensembles are designed to deal with imbalanced streams: UOB which is an extension of OB with the global oversampling scheme and UOB which is an extension of OB but with the *undersampling scheme* (Wang et al., 2015). In all ensembles 15 VFDT are used as component classifiers. All their parameters are default ones and the same as in (Brzezinski et al., 2021).

Evaluation measures: To be consistent with (Brzezinski et al., 2021) we evaluated classifiers with local accuracies of each class (Recalls) and their aggregation to G-mean[4].

Implementations of generators and examined classifiers were written in Java for the MOA data stream framework. Also all experiments were conducted in MOA command line mode similarly to ones described in (Brzezinski et al., 2021).

4. Experiments with single drifts or data difficulty factors

4.1. Imbalance between classes

The fully balanced stationary stream without any difficulty factors was quite easy to learn, where most classifiers achieved $\approx$ 0.99 G-mean and class Recalls. The average classifier performance on *stationary imbalanced streams* with minority class ratios 10%, 5%, and 3% were nearly the same — performance values rise fairly quickly up to a certain level and remain stable until the end of the stream. For instance, for imb_0.03_0.03 and the old generator G-means are (VFDT 0.9366; OOB 0.9866; UOB 0.9928; OB 0.9356) while for **N** (VFDT 0.9367; OOB 0.9519; UOB 0.9562; OB 0.9447). Even for imb_0.01_0.01 G-means are similar: 2-3% decreases for OOB and UOB, with stronger decrease for VFDT, e.g. for **O** (0.7452) and OB 0.6792. The data streams from the old generator are always slightly more difficult than for the new one.

Increasing the number of the classes is more influential for the data streams coming from the new generator, e.g. compare imb_0.03_0.03_0.03, **N** (VFDT 0.8733; OOB 0.8876; UOB 0.8896; OB 0.8801) against imb_0.03_0.03. While for the older generator the increase of the number of class is not so visible. Considering scenarios with different class imbalance ratio, such as e.g. imb_0.03_0.06_0.12. we did not observe big differences of G-mean to the configuration with the same ratios, although usually the local Recall for the second class was slightly worse.

3. Recall that our study aims at detailed investigating the impact of selected factors on the classification of multi-class imbalanced stream, not at comparing many different classifiers such as (Aguiar et al., 2022), so it is sufficient to select few representative classifiers only.
4. Due to page limits more figures and detailed results, in particular values of Recall measures are provided in an electronic appendix https://www.cs.put.poznan.pl/jstefanowski/pub/append-streams.pdf.

Comparing the classifiers we can establish the following ranking UOB $\succeq$ OOB $\succ$ OB $\approx$ VFDT. For imbalance ratios ($\geq 5\%$) the differences between them are smaller while the differences increase for the stronger imbalances. Again these difference are smaller for **N** and are more visible for **O**.

For *imbalance ratio drifts* in data streams we did not notice big differences to the static equivalents due to the values of measures. For stronger drifts (1%, 3% or 5%) a small difference in G-mean (0.01-0.03) occurs mainly for VFDT and OOB but the final values are similar or even slightly higher than for the static streams. OOB and UOB generally are not affected by these drifts. For instance for dimb_0.01_0.01; G-mean values in moments $pre \rightarrow postdrift$ are the following **O** (VFDT 0.9901 $\rightarrow$ 0.9852; OOB 0.9933 $\rightarrow$ 0.9915 ; UOB 0.9933 $\rightarrow$ 0.9940; OB 0.9914 $\rightarrow$ 0.9874) while for **N** (VFDT 0.9681 $\rightarrow$ 0.9544 ; OOB 0.9714 $\rightarrow$ 0.9602; UOB 0.9703 $\rightarrow$ 0.9606; OB 0.9704 $\rightarrow$ 0.9566). For studied configuration final values of G-means in drifting streams are slightly higher ($\geq 3\%$) than for static imbalanced stream (usually approx. 0.03) for the both generators.

A comparison with the binary imbalanced streams: The obtained results for the old generator are quite similar except classes with imbalance ratios 1%, where for multiple classes the values of measures for UOB and OOB classifiers are better than for previous binary class streams. The ranking of examined classifiers is also the same.

4.2. Borderline and rare types of examples

The G-mean results for all static configurations with different proportions of the types of examples in minority classes are presented in Table 1. The minority classes imbalance ratios are set to 10% (following the previous section it is not an influential ratio).

For static **borderline** examples one can easily notice the big decrease of G-mean for all classifiers, in particular for the older generator. Increasing the percentage of the borderline examples in the minority classes causes a very marked decrease for VFDT and OB while being smaller for OOB and UOB (but they loose 0.1 for **O**, while much more for **N**). Increasing the number of minority classes does not change these trends. Recalls of the second minority class, in the data streams from the new generator are lower than other minority classes. It may be caused that its neighborhood classes also increase their overlapping.

Borderline ratio drift: For both generators a decrease in G-mean was observed after the drift and classifiers did not recover to pre-drift levels. OB and VFDT performance is clearly worse than UOB and OOB. Some exemplary results are presented in Table 2.

For the old generator all classifiers achieve similar levels of G-Mean after the drift as for the static borderline ratio except ratio 100% where values are higher. The increase of the number of minority classes does not affect performance. For the new generator all classifiers achieve a bit higher G-mean than for the static counterpart, again except ratio 100% (which was completely deteriorated for the static streams). For the new version of the generator, an increase of the number of minority classes decreases G-means (by 0.05-0.1). Performance of classifiers OB and VFDT is clearly worse than UOB and OOB. On contrary to the imbalance ratios, now OOB is slightly better than UOB.

A comparison with the binary imbalanced streams: For multiple minority classes the decreases of the measures are higher than for binary imbalanced streams. It concerns both static and drifting scenarios. For instance, for binary class, static streams and percentage

Table 1: Comparison of the influence of various static levels of borderline or rare types of minority examples on classifier performance. G-mean is averaged over entire streams.

datastream	Old generator				New generator			
	VFDT	OOB	UOB	OB	VFDT	OOB	UOB	OB
Borderline : the percentage of example types in minority classes								
20 20	0.8600	0.9100	0.8956	0.8606	0.9558	0.9681	0.9701	0.9617
40 40	0.7533	0.8943	0.8623	0.7474	0.8357	0.8773	0.8675	0.8346
60 60	0.6978	0.8950	0.8639	0.6810	0.665	0.7692	0.7752	0.6684
80 80	0.6147	0.8922	0.8424	0.6006	0.1467	0.5367	0.5842	0.0902
100 100	0.4326	0.8954	0.8140	0.4370	0.0	0.1903	0.4834	0.0
20 20 20	0.8588	0.9071	0.8939	0.8515	0.9022	0.9069	0.9102	0.9054
40 40 40	0.7780	0.8669	0.8523	0.7566	0.7372	0.7710	0.7764	0.7370
60 60 60	0.7287	0.8554	0.8445	0.7106	0.5809	0.6732	0.6787	0.5747
80 80 80	0.6921	0.8485	0.8227	0.6841	0.1540	0.4450	0.4704	0.0949
100 100 100	0.5920	0.8384	0.7934	0.5679	0.0	0.1578	0.3737	0.0
Rare cases : the percentage of example types in minority classes								
20 20	0.8442	0.8492	0.8255	0.8296	0.8312	0.8464	0.8499	0.8374
40 40	0.6826	0.6990	0.6770	0.6855	0.6963	0.7680	0.7400	0.6949
60 60	0.5177	0.5314	0.5275	0.5131	0.5365	0.7058	0.6567	0.5275
80 80	0.2599	0.3329	0.4351	0.2559	0.3559	0.6186	0.5991	0.3276
100 100	0.0052	0.0095	0.3502	0.0	0.1551	0.5557	0.5614	0.0350
20 20 20	0.8163	0.8328	0.7882	0.8133	0.7763	0.7762	0.7781	0.7793
40 40 40	0.6639	0.6688	0.6093	0.6647	0.6287	0.6889	0.6516	0.6300
60 60 60	0.4634	0.4910	0.4723	0.4811	0.4601	0.5988	0.5426	0.4651
80 80 80	0.1999	0.2973	0.1505	0.0949	0.2789	0.4877	0.4715	0.2595
100 100 100	0.0046	0.002	0.0173	0.0	0.0985	0.3994	0.3847	0.1175

of borderline examples 20%, classifiers' performance was ≈ 0.95 while for multiple classes OOB, UOB are ≈ 0.90 and OB, VFDT are ≈ 0.85. Percentages of borderline examples closer to 100% completely deteriorate the classifiers' performance. Furthermore their reactions to the borderline drifts are also stronger with the worse recovery than for the binary streams.

Rare examples: *Stationary streams* - the influence of an increasing the percentage of rare types of minority examples is similar to static borderline examples, i.e. they decrease the performance of all classifiers although values of all measures are clearly lower; see Table 1. For higher percentages of rare examples VFDT and OB are unable to learn classes (for borderline case it appeared only for **N** and 100%) while OOB and UOB perform better (although with worse values than for similar configurations of borderline examples). For the new generator and for the percentage of rare examples at least 40% classifiers (OB, VFDT) are clearly separated from specialized methods (OOB,UOB). For the old version of generator the number of minority classes decreases G-Mean values for at least 40% rare percentage. For the new generator influence of the number of minority classes on G-Mean values increases with rising rare ratio (G-Mean value decreases by 4-15% for 3 classes

Table 2: Comparison of the influence of selected drifts of borderline or rare types of minority examples on classifier performance. G-mean is calculated in three moments of the stream.

data stream	Old generator				New generator			
moments	VFDT	OOB	UOB	OB	VFDT	OOB	UOB	OB
Borderline : the percentage of example types in minority classes								
start	0.9628	0.9826	0.9839	0.9612	0.9190	0.9191	0.9192	0.9193
d_20_20 pre	0.9347	0.9459	0.9456	0.9345	0.9185	0.9264	0.9239	0.9212
post	0.8501	0.8816	0.8880	0.8467	0.9025	0.9119	0.9052	0.9054
end	0.8528	0.9064	0.8880	0.8483	0.9065	0.9192	0.9157	0.9119
d_40_40 pre	0.8894	0.9041	0.8991	0.8882	0.8938	0.9087	0.9059	0.8982
post	0.7243	0.8269	0.8365	0.7067	0.8077	0.8515	0.8554	0.8147
end	0.7818	0.8934	0.8479	0.7651	0.8427	0.8700	0.8636	0.8476
d_60_60 pre	0.8482	0.8700	0.8780	0.9492	0.9170	0.9410	0.9454	0.9221
post	0.6273	0.7733	0.8371	0.6085	0.6956	0.8101	0.7979	0.7006
end	0.6749	0.8969	0.8797	0.6653	0.7904	0.8536	0.8532	0.7758
d_100_100 pre	0.7519	0.7982	0.8128	0.7506	0.6735	0.7731	0.7759	0.6731
post	0.4407	0.7657	0.8159	0.3942	0.0115	0.3410	0.1057	0.0087
end	0.4815	0.8906	0.8451	0.3506	0.0122	0.5103	0.5640	0.0029
Rare cases : the percentage of example types in minority classes								
start	0.9628	0.9826	0.9839	0.9612	0.9190	0.9281	0.9313	0.9294
d_20_20 pre	0.9274	0.9409	0.9412	0.9278	0.8806	0.8904	0.8864	0.8882
post	0.8246	0.8427	0.8328	0.8273	0.7868	0.7918	0.7932	0.7888
end	0.8086	0.8346	0.8275	0.8178	0.7890	0.7950	0.7966	0.7989
d_40_40 pre	0.885	0.8959	0.8956	0.8882	0.8837	0.8957	0.8941	0.8879
post	0.6913	0.7022	0.7005	0.6918	0.6955	0.7256	0.7177	0.6957
end	0.6842	0.6961	0.6967	0.6854	0.6895	0.7723	0.7755	0.6913
d_60_60 pre	0.8370	0.8469	0.8479	0.8348	0.8394	0.8553	0.8491	0.8403
post	0.5335	0.5444	0.5404	0.5336	0.5285	0.6396	0.6103	0.5300
end	0.5248	0.5385	0.5361	0.5247	0.5338	0.7058	0.7094	0.5328

scenarios). While comparing different classifiers, OOB has G-Mean values slightly higher (around 3%) than UOB. In general streams with rare examples are clearly more difficult than borderline examples and both are worse than streams affected by high imbalance only.

Local drifts of rare types of minority examples: some results are shown in Table 2 and in Figure 1. For both generators a decrease of G-Mean values is observed after the drift and classifiers do not return to pre-drift levels. In most cases there is not even a small increase in G-Mean. All classifiers after drift finally achieve similar levels of G-Mean as for static rare ratio experiment - for instance compare values of d_rare_40_40 and rare_40_40 values. For the old version of generator, the higher number of minority classes decreases G-Mean values for rare ratio after drift. For rare ratios 20%, 40%, 60%, 80% decrease is 2%, 2%, 5%, 5-11% respectively. For the new generator and for rare ratio after drift equal at least 40% classifiers (OB, VFDT) are clearly separated from specialized methods (OOB,UOB) –

they have lower G-Mean values – while for the old generator similar differences occur for the higher percentages of rare examples drifts.

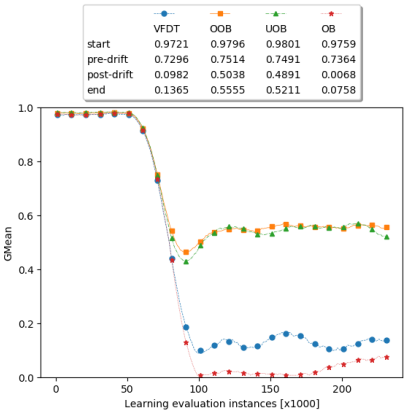

Figure 1: G-mean plot for data stream Rare drift_100_100 and N generator

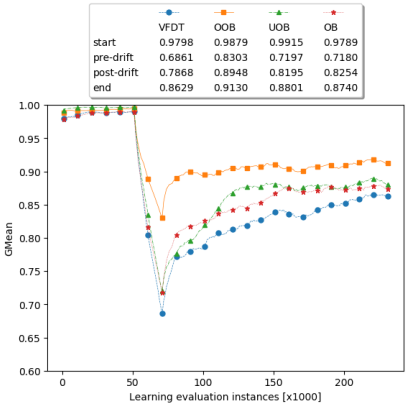

Figure 2: G-mean plot for Split_5_5 and O generator

A comparison with the binary imbalanced streams: There is some similarity between static binary and multiple class problems. For the percentage of 20% rare examples G-Mean values for binary problem are around 0.85 while for multiple minority class streams around 0.83-84. For drifts of rare examples the increasing the number of minority classes additionally decreases of G-mean values after drifts. Finally it is clearly visible that for both generators the rare example drifts are more difficult than for binary class streams.

4.3. Class composition splits and move

Experiments with minority classes moves: Following (Brzezinski et al., 2021) we model configurations of decomposing each minority class into 3, 5 and 7 sub-clusters and, then, model their moves into new positions (randomly chosen but in such a way that they not intersect). For both generators the cluster movement drift causes G-Mean decrease, although decrease is not permanent and classifiers well recover. Values for **N** generator are usually slightly higher than for **O**. Both decrease and recovery are visible in contrast to previously discussed drift scenarios (imbalance ratio, instance type). For both generators adding the third minority class causes G-Mean decrease by 0.03 – 0.05.
Experiments with minority cluster splits: The drifts are modeled as starting with single cluster shape for each minority class and then splitting it into 3, 5 or 7 non-overlapping sub-clusters. Some results are present in Table 3. For both versions of generators, the class split causes decrease of G-Mean values. What is more, classifiers do not return to pre-drift levels – see an example in Figure 2. Increasing the number of splits decreases values of G-mean after drifts for the old generator while is not so influential for the new one. The rare example drift and then borderline drift are more difficult than the split drift. For example for Bord_20_20 all classifiers have G-Mean values lower than 0.91 and for Rared_20_20 – lower than 0.86 while for all split drifts all classifiers have G-Mean values above 0.93. It was expected because rare and borderline drifts increase the number of minority instances in majority class significantly more than split drift, which leaves the overlap ratio unchanged.

Table 3: Classifiers' performance for drifts of minority classes splits into sub-clusters and one representative move drift. G-mean is calculated in four moments of the stream - start, pre-drift, post-drift and the end.

data stream	Old generator				New generator			
moments	VFDT	OOB	UOB	OB	VFDT	OOB	UOB	OB
Split each minority class into sub-clusters								
3_3 start	0.9725	0.9875	0.9876	0.9684	0.9836	0.9881	0.9921	0.9840
pre	0.7314	0.8494	0.8135	0.700	0.8836	0.9496	0.8773	0.8894
post	0.8400	0.9152	0.9016	0.8565	0.9474	0.9764	0.9608	0.9607
end	0.8936	0.9240	0.9168	0.8981	0.9660	0.9794	0.9760	0.9666
5_5 start	0.9798	0.9879	0.9915	0.9789	0.9893	0.9928	0.9939	0.9908
pre	0.6861	0.8303	0.7197	0.7180	0.8722	0.9361	0.8851	0.8817
post	0.7868	0.8948	0.8195	0.8254	0.9382	0.9691	0.9474	0.9523
end	0.8629	0.9123	0.8801	0.8740	0.9512	0.9691	0.9528	0.9583
Move sub-cluster drift								
5_5 start	0.668	0.8847	0.8523	0.6553	0.7799	0.7935	0.7803	0.7875
pre	0.7254	0.8103	0.8018	0.7368	0.8586	0.9098	0.8548	0.8633
post	0.8505	0.9510	0.8766	0.8780	0.8954	0.9150	0.8561	0.9003
end	0.9139	0.9766	0.9299	0.9525	0.9308	0.9487	0.8772	0.9332

For both generators OOB has the best G-Mean and Recall. Moreover, for this classifier decrease after drift is the lowest.

A comparison with the binary imbalanced streams: G-Mean and also Recall values after these drifts decrease and classifiers do not return to the pre-drift levels of measures. For binary problem G-Mean values decreases from 0.98 to ≈ 0.96 for all classifiers. For multiple minority class problem G-Mean values decreases from to 0.86-0.92 and classifiers have different G-Mean values. It is worth noting that for binary problem G-Mean decrease after drift is around 10-15% and it is almost the same for all classifiers. Similar observation are made for Recalls of the minority classes.

5. Experiments with complex scenarios

The interactions between data difficulty factors and drifts is examined to see how much such conjunctions additionally decrease the classifiers' performance.

Firstly we combine the imbalance ratios with all other factors. Generally they do not change the trends for these factors. Unlike the binary imbalanced streams (Brzezinski et al., 2021) (where the minority class ratios 1% – 3% amplified decreases of the measures) here only the highest ratio 1% led to very small additional decreases and did not change shapes of plots. The more interesting results are for combinations of other factors. We have chosen the split the minority classes into 5 sub-clusters (recall that 3 splits give similar results) and the drift of instance type examples of borderline or rare with two variants of percentages

20% and 60% (the second is more influential and plausible with respect to occur in streams in contrast to the higher values).

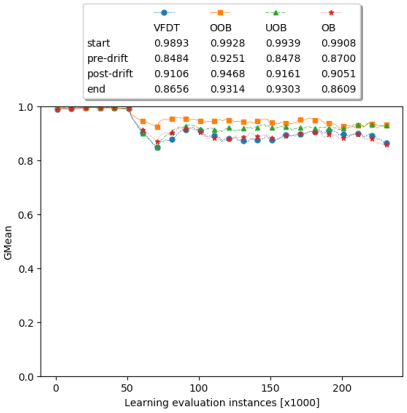

Figure 3: G-mean plot for data stream Split_5_5_Imbd_1_1_Bord_60_60 and N generator

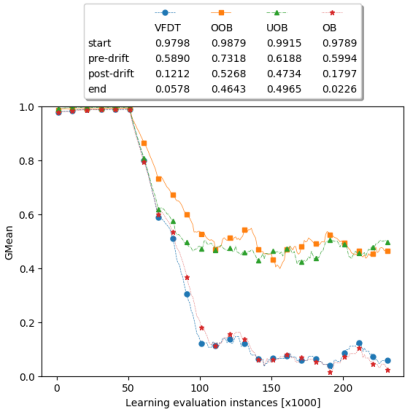

Figure 4: G-mean plot for Split_5_5_Imbd_1_1_Bord_60_60, data stream and O generator

Cluster splitting and borderline drift. For the older generator this scenario led to the lower values of G-mean than for the borderline drift alone by 0.07 – 0.18 depending on the classifier. On the other hand for the new generator differences are smaller ≈ 0.02 (Compare Figures 3 and 4). Adding the third minority class decreases of G-mean values by 0.03-0.04, while for the new generator it does not influence the results. There is also a clear difference between the better performing classifiers (UOB, OOB) and worse (VFDT,OB) for **O** generator while much less for **N**.

Comparison with binary streams: For binary problems, the split drift dominated the borderline drift in contrast to multiple minority class problem, where the result is a combination of both phenomena. G-Mean values decrease about 0.12 comparing to split experiment for binary problem. Values of G-Mean are 0.07-0.18 lower than for borderline ratio drift. Generally, the decreases of G-Mean for multiple classes are 0.05-0.10 stronger than for the binary stream.

Cluster splitting and rare drift. This combination also strongly decreases the values of measures with respect to single factors. For instance, G-Mean values after drift are lower than for rare drift by 0.03-0.08 for OOB, 0.05-0.16 for OB and VFDT and 0.09-0.20 for UOB for **O** generator while less for **N** generator (compare Figures 5 vs. 6). Moreover, the plots for classifiers are similar to the plots for rare drift, which means that for above 40% rare ratio classifiers (VFDT,OB) are clearly separated from better specialized (OOB, UOB) - see Figure 6. Finally OOB is the best classifier in this ranking.

Comparison with binary streams: the split drift seems to be dominated by the rare drift. Similarly to the binary streams, the composition of split drift with rare drift is more difficult than the split drift with borderline drift.

Then, we considered the adding the imbalanced drift 1% to these combined scenarios. These are the most difficult streams. For both generators and the variants of borderline drift it decreases G-Mean values for all classifiers, i.e. for the new generator by 0.01-0.09

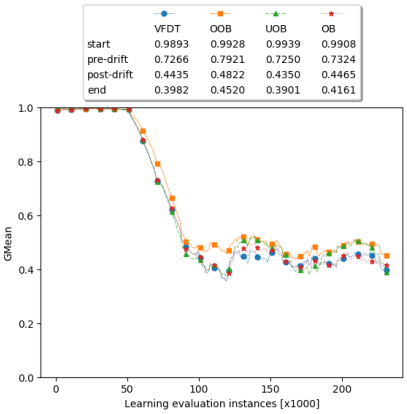

Figure 5: G-mean plot of all classifiers for Split_5_5_Imbd_1_1_Rared_60_60 data stream and N generator

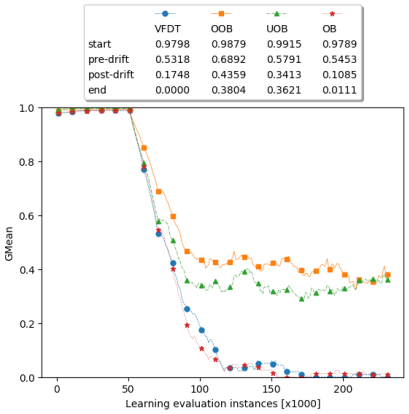

Figure 6: G-mean plot for Split_5_5_Imbd_1_1_Rared_60_60 data stream and O generator

and for the old one by: 0.05-0.26 for OOB, 0.09-020 for UOB 0.38-0.54 for OB and VFDT (which is stronger than for two combined elements). The similar amplifications of measures' decreases also occur for combining cluster splitting, rare drift and imbalance drift. Here we observed the strongest decreases for all classifiers. Moreover, an increasing of the number of minority classes decreases G-mean and Recalls (more for **O** generator).

Comparison with binary streams: The scenario with the borderline drift was not so difficult as for multiple classes. Previously UOB was the best performing classifiers while it is now overtaken by OOB. For multiple classes the stream with the rare example drift is more difficult than for binary streams and it is more clearly visible that VFDT and OB are almost unable to learn for the higher rare ratio.

To sum up our observations, generating combined factors and drifts in a multiple minority class version makes the problem more difficult than counterparts with binary classes.

6. Final remarks

The conducted experiments with multi-class imbalanced synthetic data streams allowed us to identify the most influential factors (ranking in the order of their impact): presence of rare types of minority examples, borderline examples and split of the minority class into sub-clusters. The drift of moving sub-clusters in each classes had very little impact – all classifiers well recovered after the drift to sufficiently high values. The imbalance ratios of minority classes were also not so influential. The classifiers specialized for imbalanced streams, i.e. OOB and UOB, performed really well in all of experiments with the static or drifting imbalance ratios, even for 1% which was not a case for binary streams.

The combinations of three drifts: rare or borderline types of examples, the split of the minority classes into sub-cluster and their 1% imbalances are the most difficult versions of streams. In particular the scenario with the rare examples is the most difficult one and it is more demanding than its binary class counterpart (the difference between the best classifiers for binary and the multiple minority class problem is $\approx 15\%$). Looking at figures

6 or 4 one can easily notice the high decreases of G-mean and the inability of all classifiers to recover their performance. Especially (VFDT, OB) lose the ability to recognize minority classes. Although comparing classifiers was not the aim of this work, the domination of specialized OOB and UOB ensembles in all experiments is clearly visible.

To sum up, the multiple minority class streams demonstrated to be more difficult than binary ones. Furthermore increasing the number of minority classes worsened the G-mean values ($\approx 10\%$). In particular, the recognition of the middle class for the new generator stream was weaker, which may be related to the greater overlap between the classes.

Comparing the current results to the previous ones with binary classes, one should notice a much greater importance of class overlap, i.e. borderline types of examples. In addition, the significantly lower importance of strong imbalances - which is also particularly visible for the streams from the new generator. This could inspire further research on new classifiers which could better cope with the presence of rare or borderline types of examples in imbalanced streams.

Acknowledgements: The research of Jerzy Stefanowski was partially supported by 0311/ SBAD/0726 PUT University grant.

References

Gabriel Aguiar, Bartosz Krawczyk, and Alberto Cano. A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. *arXiv preprint arXiv:2204.03719*, 2022.

Dariusz Brzezinski, Leandro L Minku, Tomasz Pewinski, Jerzy Stefanowski, and Artur Szumaczuk. The impact of data difficulty factors on classification of imbalanced and concept drifting data streams. *Knowledge and Information Systems*, 63(6):1429–1469, 2021.

Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from Imbalanced Data Sets*. Springer International Publishing, 2018.

Joao Gama. *Knowledge Discovery from Data Streams*. Chapman and Hall, 2010.

Mateusz Lango and Jerzy Stefanowski. What makes multi-class imbalanced problems difficult? an experimental study. *Expert Systems with Applications*, 199:116962, 2022.

Krystyna Napierała and Jerzy Stefanowski. Types of Minority Class Examples and Their Influence on Learning Classifiers from Imbalanced Data. *Journal of Intelligent Information Systems*, 46:563–597, 2016.

Shuo Wang, Leandro L. Minku, and Xin Yao. Resampling-based ensemble methods for online class imbalance learning. *IEEE Trans. Knowl. Data Eng.*, 27(5):1356–1368, 2015.