On the Dynamics of Classification Measures for Imbalanced and Streaming Data

Dariusz Brzezinski, *Member, IEEE,* Jerzy Stefanowski, Robert Susmaga, and Izabela Szczęch

(Special Issue on Recent Advances in Theory, Methodology and Applications of Imbalanced Learning)

***Abstract*—As each imbalanced classification problem comes with its own set of challenges, the measure used to evaluate classifiers must be individually selected. To help researchers make this decision in an informed manner, experimental and theoretical investigations compare general properties of measures. However, existing studies do not analyze changes in measure behavior imposed by different imbalance ratios. Moreover, several characteristics of imbalanced data streams, such as the effect of dynamically changing class proportions, have not been thoroughly investigated from the perspective of different metrics. In this paper, we study measure dynamics by analyzing changes of measure values, distributions, and gradients with diverging class proportions. For this purpose, we visualize measure probability mass functions and gradients. Additionally, we put forward a histogram-based normalization method that provides a unified, probabilistic interpretation of any measure over datasets with different class distributions. The results of analyzing eight popular classification measures show that the effect class proportions have on each measure is different, and should be taken into account when evaluating classifiers. Apart from highlighting imbalance-related properties of each measure, our study shows a direct connection between class ratio changes and certain types of concept drift, which could be influential in designing new types of classifiers and drift detectors for imbalanced data streams.**

***Index Terms*—classification measures, class imbalance, data streams, concept drift, measure histograms, measure gradients**

I. Introduction

PERFORMANCE of most classifiers can be considerably deteriorated when they are learned from imbalanced data. Over the last years, the problem of improving classifier performance on such skewed data has received much interest, resulting in several proposals of specialized data preprocessing methods and classifier modifications [1], [2]. Nevertheless, class imbalance is still treated as a challenging problem, especially when co-occurring with other data difficulty factors such as small dataset size, high feature dimensionality, or complex instance distributions [3], [4].

One of the important aspects of tackling imbalanced data is the selection of an appropriate classification performance measure. Since standard evaluation measures, such as *accuracy*, tend to focus on recognizing all target classes, several other measures addressing class imbalance have been proposed [5]. The number of such dedicated measures, commonly defined on the basis of confusion matrices, is relatively high. Each of them may represent different aspects of classification predictions and may lead to contrasting interpretations. Therefore, the selection of the right metric for a particular task requires careful thought. Unfortunately, such a choice is often driven by the measure's popularity rather than resulting from a thorough discussion of its properties. Moreover, comprehensive studies analyzing measure properties are rare [6]–[10] and focus on the measure as a whole, omitting analyses of changes of measure values imposed by different class proportions. In our opinion, questions such as: How may the measure's values change with respect to various class proportions?, How should one interpret a particular measure value for a given class proportion?, Is it potentially easy to improve a measure's value by modifying class distributions?, are still worth investigating.

Answering these questions would facilitate the proper interpretation of measure behavior for different class imbalance ratios. For instance, a researcher carrying out experiments on several datasets characterized by various imbalance ratios should be able to truly understand measure values for the given data. F_1*-score* equal to 0.7 may convey different amounts of information for different class distributions. Moreover, this value may be easier to improve upon for some class proportions than others. Finally, the particular combinations of predictions in the confusion matrix may infer distinct directions of the fastest changes of the measure's value (gradients). This, in turn, may relate to specialized preprocessing methods and classifier modifications for imbalanced data.

Changes in measure behavior can constitute a challenge not only for binary imbalanced data, but also for multi-class [11], [12] and multi-label [13] scenarios, including specific problems such as the concurrence of minority and majority classes in the same instance [14]. However, the aforementioned analyses are particularly important for classification of concept-drifting data streams, i.e., sequences of continuously generated data items with class and feature distributions changing (drifting) over time [15]. In imbalanced streams, one of the aspects that can drift is the class imbalance ratio. This in turn may lead to problems in interpreting classifier performance over time. Moreover, in such streams the role of classes may change, i.e., the majority class may become the minority one. These phenomena cause difficulties for developing re-sampling-based stream ensembles [16]. Even more so, the presented challenges are critical for designing drift detectors [17], which identify changes in the stream by tracking the dynamics of classifier predictions. Current solutions exploit measures such as *recall* or *G-mean* [18], however, the principle differences between their performance are still to be explained. Therefore, the appropriate interpretation of measure values and their dynamics in response to fast evolving class distributions is of particular interest to the field of data stream mining.

The authors are with the Institute of Computing Science, Poznan University of Technology, ul. Piotrowo 2, 60–965 Poznan, Poland, e-mail: dariusz.brzezinski@cs.put.poznan.pl.

Manuscript received -; revised -.

In this paper, we consider the above issues in interpreting *measure dynamics*, i.e., changes of measure values, distributions, and gradients with diverging class proportions. To analyze such dynamics we propose a histogram-based approach, which illustrates frequencies of possible values aggregated over all combinations of predictions in confusion matrices. Furthermore, we visualize changes in these histograms for different class proportions. Contrary to approaches focused on changes in classifier performance, such as cost-curves [19], the presented approach is independent of any particular classifier or dataset and, therefore, corresponds to general measure properties. Additionally, we put forward and experimentally validate a histogram-based normalization method that allows to unify measure interpretation over datasets with different class distributions. In an experiment involving seven classifiers and 12 benchmark datasets, we show how the proposed normalization method can help set a common ground for comparing multiple models on multiple problems and re-interpret which datasets offer most space for classifier improvement. Finally, to extend the knowledge on the dynamics of measure values we demonstrate how to exploit measure gradients using barycentric visualization [20] and how they relate to changes occurring in imbalanced streams. All these analyses are performed on eight popular classification measures. To the best of our knowledge, the presented approaches have not been considered in previous works on classifier evaluation measures.

The paper is organized as follows. In Section II, we discuss related works on classification measures for imbalanced data. In Section III we analyze measure behavior on the basis of their probability mass functions visualized as histograms. In particular, Section III-A discusses the dynamics of measure values with respect to class proportions and Section III-B presents a new method for measure normalization. Additionally, in Section IV we interpret measure gradients in the context of class imbalance and concept drift. Finally, Section VI concludes the paper and draws lines of future research.

II. Related Works

Good recognition of the minority class is a key requirement for most imbalanced classification problems, however, its trade off with predictions of other classes can be addressed in different ways [6], [9], [21], [22]. Although several measures have been already considered, there is no single measure that is the best in all imbalanced problems and its choice for a given dataset is not an obvious task.

In this paper, we focus on analyzing and visualizing measures that evaluate crisp classifier predictions based on the binary confusion matrix. Therefore, measures calculated using probability estimates and label rankings, such as for example Brier score or the area under the ROC curve, are out of the scope of this study. We note, however, that the tackled subject is relevant not only to binary classification problems, but may also apply to muti-class and multi-label scenarios. Finally, we would like to underline that this study should not be confused with analyses and visualizations of *classifier* performance. Our main intention is to study general properties of *measures* rather than visualize the predictive performance of a classifier on a given dataset. To achieve this goal, we dissect the effect of class imbalance on performance measures, regardless of any particular classifier, dataset, feature dimensionality or instance distribution.

A. Classification Measures for Imbalanced Data

The considered measures are functions of predictions represented in the confusion matrix for two-class problems (Table I), where the minority class is typically referred to as the positive class (P) and the majority class is referred as the negative one (N). The TP (*True Positive*) and TN (*True Negative*) entries denote the number of examples classified correctly by the classifier as positive and negative, while the FN (*False Negative*) and FP (*False Positive*) indicate the number of misclassified positive and negative examples. Additionally, $TP + TN$ is denoted as T, while $FN + FP$ is denoted as F.

TABLE I: Confusion matrix for two-class classification

Actual \ Predicted	Positive	Negative	total
Positive	TP	FN	P
Negative	FP	TN	N
total	$\widehat{P}$	$\widehat{N}$	n

For further analysis, we selected eight measures which are commonly applied in experimental studies [5]. *Precision*, *recall* (*sensitivity*), and *specificity*, are the most popular single-class measures, out of which we will focus on *precision* and *recall*. We also chose the most commonly used aggregations of single-class measures: F_1*-score*, the harmonic mean of *precision* and *recall*; *G-mean*, the geometric mean of *sensitivity* and *specificity*; and *balanced accuracy*, the arithmetic mean of *sensitivity* and *specificity*. The domains of all these measures are between 0 and 1, where 1 is the preferred value. Furthermore, *Matthews Correlation Coefficient* (*MCC*), strongly recommended in [9] and [22], is a measure expressing correlation between the actual and predicted classification, which returns values between -1 (total disagreement) and $+1$ (perfect agreement). Additionally, we analyze the *Kappa* statistic [5], which corrects accuracy for chance predictions, strongly related to class imbalance. *Kappa* can achieve values from -1 to $+1$, where zero means that the classifier is no better than a chance prediction and values above/below zero indicate how much better/worse the predictions are. Finally, to complement the analysis of measures for imbalanced data, we also take into account classification accuracy, which is the basic reference measure for any classification task. The definitions of all the considered measures are the following:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$balanced\ accuracy = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right) \tag{2}$$

$$Kappa = \frac{accuracy - \frac{1}{n}\left(\frac{P \cdot \widehat{P}}{n} + \frac{N \cdot \widehat{N}}{n}\right)}{1 - \frac{1}{n}\left(\frac{P \cdot \widehat{P}}{n} + \frac{N \cdot \widehat{N}}{n}\right)} \tag{3}$$

$$G\text{-}mean = \sqrt{\frac{TP}{TP+FN} \cdot \frac{TN}{TN+FP}} \quad (4)$$

$$F_1\text{-}score = 2 \cdot \frac{recall \cdot precision}{precision + recall} \quad (5)$$

$$precision = \frac{TP}{TP+FP} \quad (6)$$

$$recall = \frac{TP}{TP+FN} \quad (7)$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{\widehat{P} \cdot P \cdot N \cdot \widehat{N}}} \quad (8)$$

It is worth noting that most of the above measures have been used as a basis for evaluating and designing drift detectors for imbalanced streams [18], [23], [24].

B. Measure Analyses and Visualizations

The measures presented in the previous section were compared in many surveys [6]–[9], however, usually with the aim of discussing the main differences in their definitions. Additionally, the F_1-*score* was thoroughly analyzed by Powers [25] who claimed that some of its properties, such as focusing only on the minority class and assuming that actual and predicted distributions are identical, may be critical flaws. Another theoretical study showed that aggregating *sensitivity* and *specificity* presented more suitable behavior than aggregating *precision* and *recall* [6]. Nevertheless, all of the mentioned theoretical analyses did not focus on the changes of measure properties with diverging class proportions.

Apart from theoretical studies, other related works concern visual-based analyses of measures. One of such works discusses measure visualizations in 3D ROC space [26], where the author mentions *skew invariance* as one of the properties a measure may possess. Skew invariance is indeed a property related to changing class proportions, yet one that does not quantify or categorize changes in measures for consecutive class ratios. A more recent study [10] introduces visualizations in barycentric space and puts forward ten general properties with an attempt to facilitate measure selection for imbalanced data. In parts of this paper, we also make use of the barycentric space, but visualize measure gradients instead of measure values and focus on measure dynamics, normalization, and applications to stream mining. Moreover, the properties proposed in [10] concentrate on symmetries, minima, maxima, and undefined values, which do not directly describe measure dynamics with respect to class proportions.

We re-iterate that, although related, this study should not be confused with visualizations of classifier performance, e.g., using ROC graphs [27], precision-recall curves [28], or other attempts to graphically present experimental comparisons of classifiers. The works of Curuana *et al.* [29] and Alaíz-Rodríguez *et al.* [30], use multidimensional scaling to present the results of several classifiers on multiple datasets according to more than one measure. This can be considered somewhat related to the histogram-based measure normalization proposed in this paper, however, our normalization focuses on a single measure, outputs a numerical value, and is independent of the classifier and particular dataset. Finally, contrary to methods presented in this study, cost-curves [19], [31] focus on classifier dynamics on a given (static) dataset, rather than on the general dynamics of a performance measure, regardless of any concrete classifier or dataset.

III. Histogram-based measure properties

The goal of this study is to analyze the behavior of measure values for varying imbalance ratios. These values are functions of confusion matrices, which correspond to outcomes of classification on experimental data. If one interprets the training data as a result of a random process, one can give a probabilistic interpretation to classification measures. In particular, measures based on confusion matrices can be considered as discrete random variables, which map a confusion matrix to a numerical value. Discrete random variables are commonly described by their *probability mass functions* (*pmfs*), which are functions that give the probability that a discrete random variable is exactly equal to some value [32]. Probability mass functions are often displayed as histograms, where the x-axis represents measure values and the y-axis the probability of achieving a given value. We will use such visualizations to analyze the considered classification measures.

In this section, we study the dynamics of eight classification measures (1)–(8) by means of their probability mass functions depicted as histograms. In our analysis, we abstract from concrete classifiers or datasets, and, therefore, assume that each possible confusion matrix is equally probable. Consequently, we generate all possible confusion matrices for a dataset size n with class proportion $ir = P{:}N$, and calculate the measure's value for each matrix. Using the calculated measure values, we analyze the *pmf*-based histograms of each measure for varying class imbalance ratios.

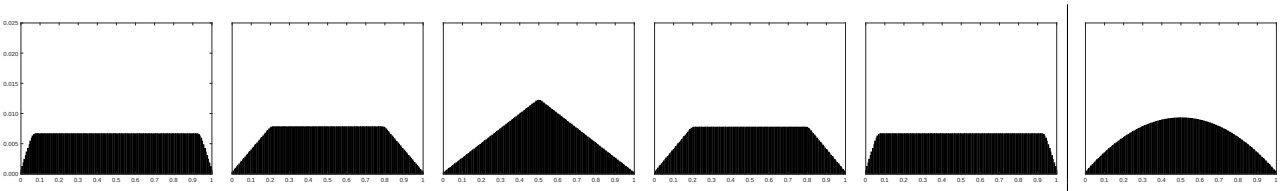
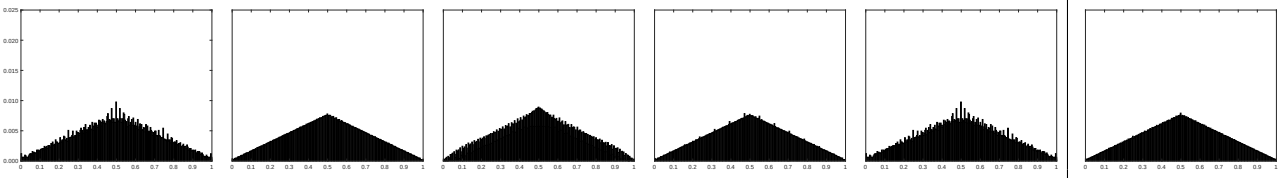
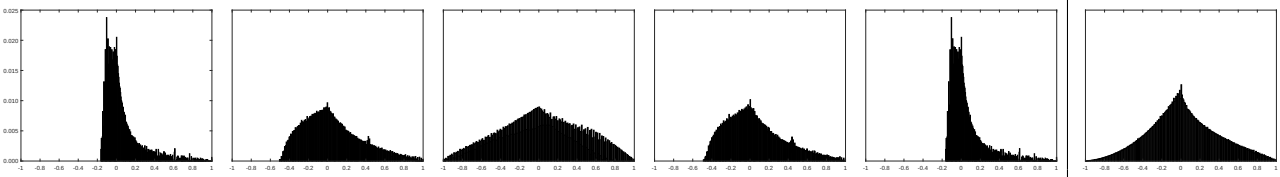
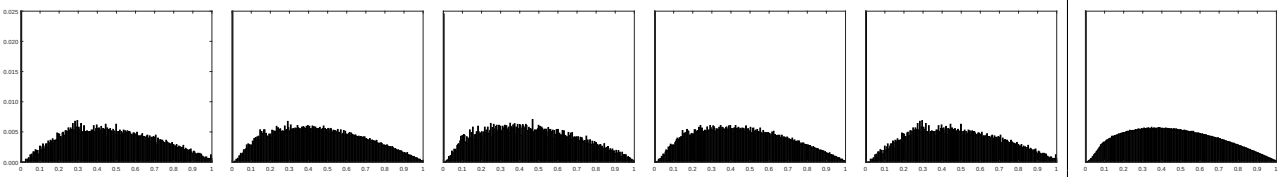
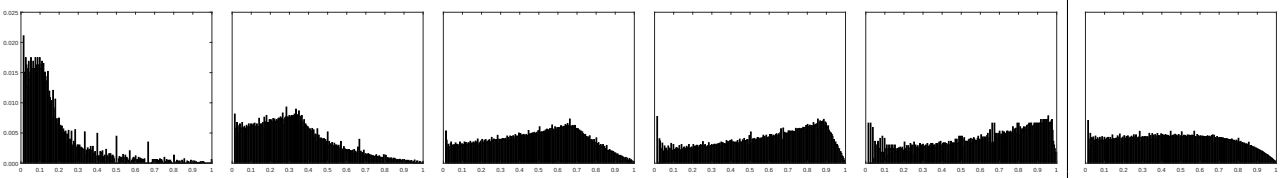
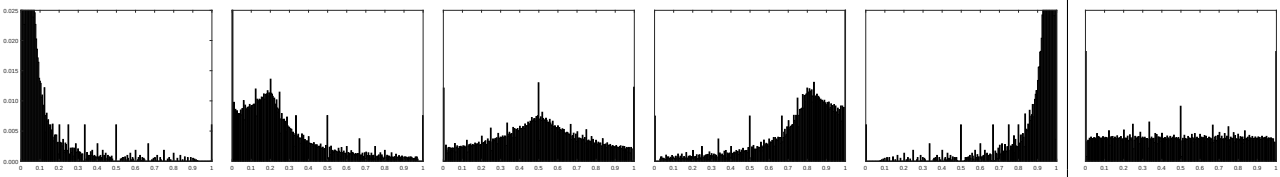
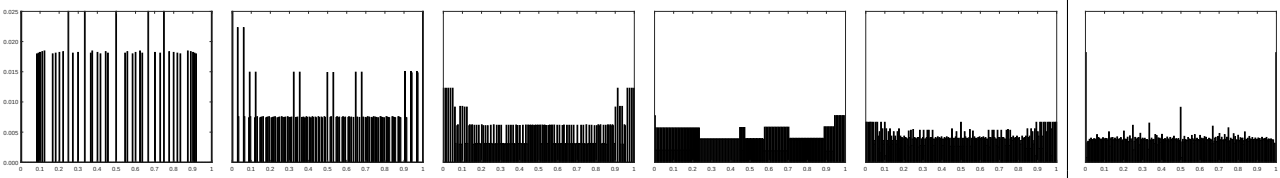
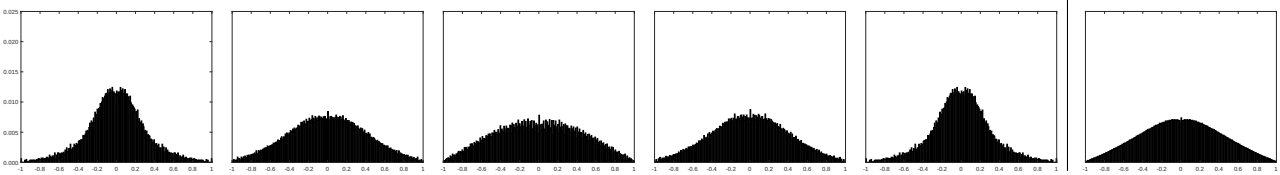
In our visualizations, we use $n = 160$ with $ir \in \{1{:}15, 1{:}4, 1{:}1, 4{:}1, 15{:}1\}$. Class proportions ir were selected to represent class balance (1:1), low imbalance (1:4/4:1), and moderate imbalance (1:15/15:1) [33], and span a range similar to that of real-world imbalanced datasets used in our experiments (Section III-B). The dataset size n was chosen to correspond to the selected class proportions and to ensure that for high imbalance (1:15/15:1) only few examples of the minority class are available. To complement the analysis, we also demonstrate histograms incorporating all possible class proportions for a given n. The resulting figures are presented in Table II. Additional visualizations for higher imbalance ratios $ir \in \{1{:}159, 1{:}79, 1{:}31\}$ and a larger dataset $n = 16000$ are available in the online supplementary material (Appendix B).

Based on the presented histograms, in Section III-A we analyze how the measures' *pmfs* change with varying imbalance ratios. Next, in Section III-B we propose a histogram-based measure normalization method.

A. Measure Dynamics Based on Imbalance-wise Distributions

The visual analysis of histograms gathered in Table II and Supplementary Tables B.1–B.2 allows us to identify differences between measure behavior occurring when the imbalance ratio changes. In particular, *accuracy* starts with a triangular distribution for the 1:1 class ratio and tends towards a uniform distribution when the imbalance in data

TABLE II: Histograms of values of eight classification measures. The x-axis spans between the minimum and maximum of each measure and is divided into 256 bins of equal width. The y-axis shows the probability of obtaining a given measure value. The y-axis was set up equally for all the measures, introducing minor clipping for *G-mean* (large number of zeros) and *Precision* (large number of low/high values for 1:15/15:1 class ratios).

Measure	1:15 class ratio	1:4 class ratio	1:1 class ratio	4:1 class ratio	15:1 class ratio	All class ratios
Accuracy						
Balanced accuracy						
Kappa						
G-mean						
F_1-score						
Precision						
Recall						
MCC						

increases. *Balanced accuracy* and *G-mean* keep the same shape of the histogram regardless of the class proportion. Observe that even the high number of zeros remain the same for all imbalance ratios of *G-mean*. On the other hand, *Kappa* is a measure characterized by an almost triangular distribution when the classes are balanced, but changes considerably with diverging class proportions. Increasing imbalance results in a compression of the left-hand side of *Kappa's* histograms corresponding to low measure values, keeping, however, the right-hand side almost unaffected. This means, that the range of possible values changes with class proportions, making it increasingly more difficult to obtain high values. Next, F_1-*score* and *precision* can be grouped together as similarly behaving, because their histograms' mass gradually shifts towards the right-hand side (high measure values) as the number of positive examples increases (compare, e.g., class ratios for 1:15 and 15:1). On the other hand, *recall* does not change the shape of its histograms, which resemble a uniform distribution, but their resolution is heavily affected by the number of positive examples. Notice how few measure values are obtained for the 1:15 class ratio, compared to 15:1. In fact, the number of different measure values is in direct proportion to the number of positive examples in the data (see also Supplementary Table B1–B.2). Finally, the histograms of *MCC* resemble a Gaussian-like distribution with $\mu = 0$. Increasing class imbalance results in decreasing the distribution's width (standard deviation).

The shapes of the histograms portray the probability mass function of the analyzed measures. Therefore, the presented plots depict probabilities of obtaining certain measure values for a given class proportion. Since the histograms differ considerably, it follows that value distributions differ between and within measures for different imbalance ratios. In particular, the histograms often contain regions where certain values are underrepresented (harder to obtain). For example, for $ir = 1{:}15$ it is much more difficult to achieve an F_1-*score* higher than 0.90, whereas that same value is fairly common when $ir = 15{:}1$. This observation can have a direct impact on classifier evaluation in scenarios where the imbalance ratio changes dynamically. Such situations can occur, for example, in concept-drifting data streams, where class definitions and their proportions can fluctuate with time. A common task in data stream mining that is connected with classifier evaluation is the detection of changes in class definitions by using drift detectors.

The basic task of any drift detector is to signal a significant change (drift) in the incoming data (concept). Drift detectors are often implemented using statistical tests based on sequential analysis and process control charts [17]. If such a test checks whether a selected performance measure is significantly different from its previous values, the probability of obtaining a given value, depicted in the analyzed histograms, sheds light on how "smoothly" a detector works for a given imbalance ratio. For example, a drift detector monitoring the F_1-*score* of a classifier characterized by F_1-*score* $= 0.90$ will work smoothly when the number of positive examples is large, e.g., for a 15:1 class ratio presented Table II. This is because for such a class ratio the number of confusion matrices producing F_1-*score* values around 0.90 is fairly large, thus, making the chance of a sudden value change rather small.

Moreover, if we assume that class proportions in the data stream can change with time [24], the awareness of the differences in measure distributions becomes even more important. If a measure's histogram changes with the imbalance ratio, then so does the performance of a drift detector. The detector monitoring the classifier's F_1-*score* that worked well when the number of positive examples was large will potentially suffer from more false alarms when the number of positive examples decreases, e.g. to a 1:15 class ratio presented in Table II.

Finally, apart from the shape of the histogram its resolution can also be influenced by class proportions. This effect can be easily observed for *recall*, which obtains very few different values when the number of positive examples is low. As a result, a drift detector based on *recall* [18] has to cope with sudden jumps of the measure's values when the number of positive examples is very low.

To experimentally validate the above observations, we performed an experiment involving a drift detector monitoring the eight considered measures. Our hypothesis was that on a static stream with no concept changes a drift detector should produce *fewer* false alarms (incorrect detections) when the value it oversees corresponds to a dense region in the pmf. In other words, we expect the drift detector to be less susceptible to noise when many confusion matrices produce similar measure values. To verify this hypothesis, we used the PH test drift detector [23] to monitor the changes in classifier performance according to measures (1)–(8). To oversee the metrics, we used a window size $w = 100$ and default PH test parameters implemented in the MOA stream testing environment [34]. The experiments monitored the performance of a Hoeffding Tree classifier on streams generated using the Agrawal generator [34].[1] The streams had a constant imbalance ratio and did not contain any concept drift. We used the Agrawal generator since, as our previous study has shown [35], the PH test is susceptible to false alarms on this dataset. Tables III and IV present the mean measure values and number of false alarms for the five analyzed imbalance ratios, respectively. The results refer to prequentially calculated [23] means and false alarm ratios over 100 streams generated with different random seeds.

TABLE III: Prequentially calculated mean measure values from 100 repetitions of the drift detector experiment; standard deviations are given in parentheses.

Measure	1:15	1:4	1:1	4:1	15:1
Accuracy	0.93(.01)	0.86(.03)	0.87(.03)	0.93(.02)	0.97(.02)
Bal. acc.	0.53(.08)	0.77(.08)	0.87(.03)	0.83(.04)	0.78(.10)
Kappa	0.09(.18)	0.53(.13)	0.74(.06)	0.73(.07)	0.66(.18)
G-mean	0.14(.22)	0.74(.13)	0.87(.03)	0.81(.05)	0.73(.14)
F_1-score	0.09(.19)	0.62(.13)	0.88(.03)	0.96(.01)	0.98(.01)
Precision	0.28(.37)	0.66(.09)	0.83(.04)	0.92(.02)	0.97(.01)
Recall	0.07(.16)	0.63(.17)	0.94(.04)	0.99(.03)	1.00(.01)
MCC	0.11(.19)	0.54(.11)	0.75(.06)	0.75(.07)	0.70(.15)

The results confirm our hypothesis and directly relate to observations made during the histogram analysis. The

[1]Source codes, datasets and reproducible scripts for all experiments are available at: https://github.com/dabrze/measure_dynamics

TABLE IV: Ratio of false alarms within 100 runs of the PH test monitoring the eight considered measures.

Measure	1:15	1:4	1:1	4:1	15:1
Accuracy	0.00	0.12	0.27	0.03	0.00
Bal. acc.	0.09	0.11	0.26	0.27	0.25
Kappa	0.68	0.44	0.50	0.43	0.49
G-mean	0.91	0.59	0.24	0.33	0.47
F_1-score	0.82	0.50	0.24	0.01	0.00
Precision	0.94	0.80	0.50	0.11	0.00
Recall	0.70	0.44	0.27	0.04	0.01
MCC	0.76	0.49	0.48	0.40	0.50

detector monitoring *accuracy*, produces more false alarms when the classes are more balanced. This is in accordance with histograms presented in Table II, where high values of accuracy ($\sim$0.9) are underrepresented for the balanced distribution compared to class imbalance. On the other hand, measures that maintain similar histogram shapes through varying imbalance ratios (*balanced accuracy*, *Kappa*, *G-mean*, *MCC*), also maintain similar levels of false alarms for the tested ratios. The slightly higher number of false alarms when $ir = 1{:}15$ is due to the fact that for this class proportion the values of the aforementioned measures where much lower, placing the analysis in a different part of the histograms. Finally, the remaining measures have false alarm rates directly proportional to the number of positives in the data stream. This relates to the moving mass of *F_1-score* and *precision*, and the resolution-effect of *recall*.

B. Histogram-based Measure Normalization

The observed differences in histogram shapes suggest that the measure's value should be interpreted differently depending on the class proportion. To address this issue, we propose a normalization method that takes into account the class imbalance in a given dataset and standardizes measure values according to their frequencies. The resulting normalized measure values gain a probabilistic interpretation that remains true for varying imbalance ratios. Therefore, one will be able to compare normalized measure values between datasets with different class proportions, and potentially display different measure values on the same scale.

Our approach is based on *cumulative distribution functions* (*cdfs*). A *cdf* of a discrete random variable x is defined as:

$$F_X(x) = P(X \leq x) = \sum_{t \leq x} pmf_X(t) \tag{9}$$

where the right-hand side represents the probability that the random variable X takes on a value less than or equal to x, which can be computed by summing the values of X's probability mass function $pmf_X()$ up to x. Since the histograms presented in Table II are based on the measures' *pmfs*, the *cdf* of each measure X can be visually interpreted as the proportion of the histogram (proportion of confusion matrices) with a value less than or equal to x.

Therefore, we propose the following normalization method: For a given measure M on a dataset with a class ratio ir, the measure's value x is normalized according to:

$$N(M, ir, x) = \sum_{t \leq x} pmf_{M,ir}(t) \tag{10}$$

where $pmf_{M,ir}()$ is the probability mass function of measure M for proportion ir. The resulting value of the normalization is always between 0 and 1, and represents the probability that measure M takes on a value less than or equal to x. Thus, the higher $N(M)$, the smaller the chance of improving a classifier, in terms of M, for a given dataset.

We will illustrate the normalization method on a simple example. Let us assume we have a dataset with $n = 160$ examples, where the number of positives and negatives are $P = 150$ and $N = 10$, respectively. Next, let us assume that a classifier achieved a value of *precision* $x = 0.9$. For any given P and N, the number of all possible confusion matrices is $c = (N{+}1)(P{+}1)$. In our example, $c = (150{+}1)(10{+}1) = 1661$. The number of confusion matrices for which the value is lower than x depends on the measure and class proportions, and has to be computed by applying the measure to the confusion matrices. In our case, by applying (6) to c confusion matrices we get 506 matrices for which *precision* is less than or equal to 0.9. Therefore, $N(precision, 150{:}10, 0.9) = 506/1661 \approx 0.3$. If one performed a similar operation for $P = 10$ and $N = 150$, the result would be $N(precision, 10{:}150, 0.9) = 1650/1661 \approx 0.99$. As the example shows, it is much more difficult to improve a classifier that achieved 0.9 *precision* on a dataset with few positive examples than it is to improve that result when the number of positives is high (even though in both cases *precision* was the same).

To verify how the normalization method works for different measures on real data, we performed an experiment on 12 benchmark imbalanced datasets from the UCI repository [36] (Table V). The datasets were selected to represent various sizes, various imbalance ratios, and feature characteristics [4].

TABLE V: Characteristic of real-world benchmark datasets used to analyze the proposed histogram-based measure normalization.

Dataset	Examples	Features	Imbalance ratio	Minority class
arcene	200	10000	$\sim$1:1	positive
breast-w	699	9	$\sim$1:2	malignant
colon	62	2000	$\sim$1:2	1
credit-g	1000	61	$\sim$1:2	bad
ecoli	336	7	$\sim$1:9	imU
glass	213	9	$\sim$1:12	v-float
ionosphere	350	34	$\sim$1:2	bad
micromass	571	1300	$\sim$1:10	AUG.AEX
new-thyroid	214	5	$\sim$1:5	hyper
solar-flare	1066	10	$\sim$1:5	F
transfusion	747	4	$\sim$1:3	yes
yeast	1484	8	$\sim$1:32	ME2

Using the selected datasets, we performed 10 repetitions of stratified 10-fold cross-validation [5] to evaluate seven types of learning algorithms, chosen for their diversity: k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Naive Bayes (NB), Decision Tree (CART), Logistic Regression (LR), Random Forest (RF), and Gradient Boosting Machines (GBM). Since our goal is only to illustrate the effect of

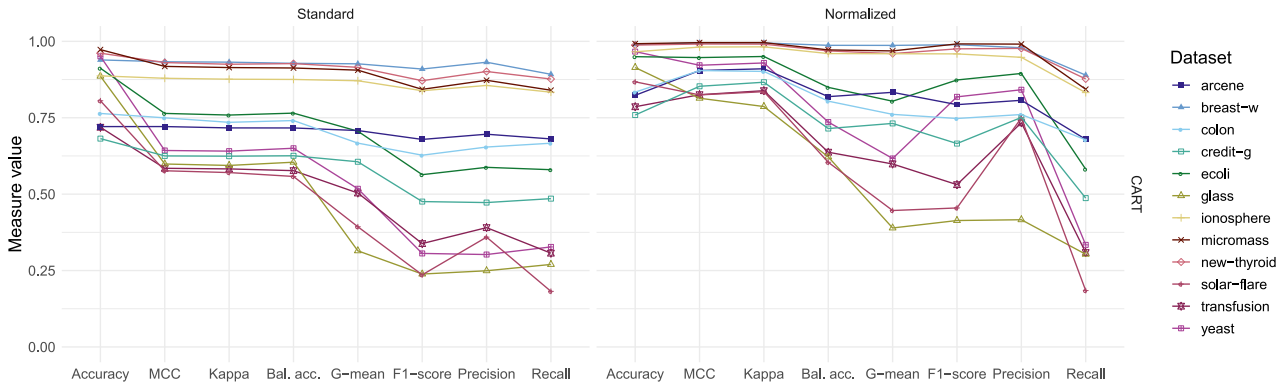


Fig. 1: Comparison of CART performance on 12 imbalanced datasets using standard measures M (left panel) and their normalized counterparts $N(M)$ (right panel). Mean results from 10 repetitions of stratified 10-fold cross-validation. The values of standard MCC and Kappa were 0–1 scaled to obtain a common scale for all measures; no scaling was necessary for normalized measures.

measure normalization, we left the classifiers with default parameters in their Python implementation in the scikit-learn library [37]. Figure 1 presents a parallel coordinate plots comparing CART results obtained for the analyzed measures (left panel) and their normalized counterparts (right panel). Detailed tabular results and plots for the remaining classifiers are available in Supplementary Tables A.1–A.2 and Figures A.1–A.6, respectively.

As the results show, the effect of the normalization depends on the measure being normalized and the class proportion in a dataset. For example, the evaluation of CART (Figure 1, left panel) suggests that with 0.30 *precision* `yeast` is the dataset with second lowest performance. However, the normalized value of *precision* is 0.84 (Figure 1, right panel), which means that it is in fact a dataset for which CART obtained a confusion matrix better than 84% of all possible confusion matrices. When comparing this result with other values of normalized *precision*, we can see that this makes `yeast` one of the datasets on which CART performs well. Analogously, normalized F_1*-score* uncovers fairly good performance of CART on datasets such as `yeast`, `ecoli`, `credit-g`, not evident when analyzing standard F_1*-score*. A similar, yet milder effect can be observed when comparing standard and normalized versions of *MCC* and *Kappa*. On the other hand, since the histograms of *recall* resemble a uniform distribution, normalization has almost no effect on its value. Finally, normalized versions of *accuracy*, *G-mean*, and *balanced accuracy* tend to promote good recognitions for balanced datasets, a phenomenon which can be noticed when comparing the results on `ionosphere` with those obtained on `ecoli`. These examples illustrate that the proposed normalization method can give new perspective to results obtained for each measure.

The effects of normalization were clearly noticeable for all 7 analyzed classifiers (Figure 1, Supplementary Figures A.1–A.6). The comparison of parallel coordinate plots of different classifiers shows how the normalization method takes into account both the datasets imbalance ratio and the obtained measure value. For example, NB is the only classifier that succeeds at classifying `glass` better than other datasets, according to the normalized measure values. Since this reinterpretation of results on `glass` is not consistent for all the classifiers, it clearly shows that the proposed normalization is not a simple 0–1 scaling procedure but a method that takes into account the measures characteristic, imbalance ratio, and concrete confusion matrix. Additionally, experiments on synthetic data with varying class proportions and numbers of features (supplementary material, Appendix C) show that, although some classifiers perform better than others for certain data, measure values are indeed affected by the imbalance ratio, regardless of other data characteristics.

We note that the proposed normalization method will not affect model selection based on a single measure using a single dataset. On a single dataset all classifier results for a single measure are normalized according to the same pmf, therefore, while the interpretation of classifier performances will change, their ordering will not. However, since the normalization gives a probabilistic interpretation to each measure, it allows to compare several measures on the same scale on several datasets. This, in turn, could be used to visually and numerically compare multiple classifiers on multiple datasets, in a fashion similar to that proposed by Curuana *et al.* [29] and Alaíz-Rodríguez *et al.* [30].

The histograms and normalization method discussed in this section highlighted the dynamics of measures related to varying class proportions. In the following section, we study the speed of measure changes by analyzing their gradients and show the connection between gradients for different class ratios and concept drift.

IV. Measure Gradients

To analyze measure gradients we use a recently proposed visualization technique for analyzing classification measures in a barycentric space [10]. In Section IV-A we recall the basics of this technique, and extend it to study measure gradients in Section IV-B.

A. Barycentric Measure Visualization

The *barycentric coordinate system* is a coordinate system in which point locations are specified relatively to hyper-sides

of a simplex. A 4D barycentric coordinate system is a tetrahedron, where each dimension may be thought of as represented by one of the four vertices. Choosing vectors that represent TP, FP, FN, TN as vertices of a regular tetrahedron in a 3D space, one arrives at a barycentric coordinate system depicted in Fig. 2.

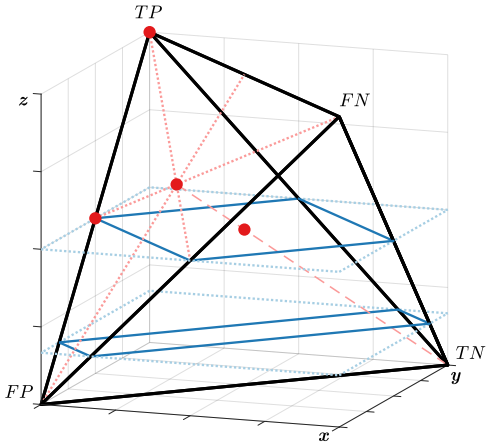


Fig. 2: A skeleton visualization of the tetrahedron with four (red) points corresponding to four exemplary confusion matrices and two rectangular (blue) cross-sections corresponding to two class proportions.

In this system, every confusion matrix $\left[\begin{smallmatrix} TP & FN \\ FP & TN \end{smallmatrix}\right]$ is represented as a point of the tetrahedron. Let us illustrate this fact with a few examples. Figure 2 shows a skeleton of a tetrahedron with four exemplary points:

- one located in vertex TP, which represents $\left[\begin{smallmatrix} n & 0 \\ 0 & 0 \end{smallmatrix}\right]$,
- one located in the middle of edge TP–FP, which represents $\left[\begin{smallmatrix} n/2 & 0 \\ n/2 & 0 \end{smallmatrix}\right]$,
- one located in the middle of face △TP–FP–FN, which represents $\left[\begin{smallmatrix} n/3 & n/3 \\ n/3 & 0 \end{smallmatrix}\right]$,
- one located in the middle of the tetrahedron, which represents $\left[\begin{smallmatrix} n/4 & n/4 \\ n/4 & n/4 \end{smallmatrix}\right]$.

One way of understanding this representation is to imagine a point in the tetrahedron as the center of mass of the examples in a confusion matrix. If all n examples are true positives, then the entire mass of the predictions is at TP and the point coincides with vertex TP. If all examples are false negatives, the point lies on vertex FN, etc. Points corresponding to all possible confusion matrices for a given class ratio are represented as rectangular cross-sections in the tetrahedron. Figure 2 depicts two cross-sections: one for class balance (middle of the tetrahedron) and one for a 1:5 imbalance ratio (lower part of the tetrahedron).

Using the barycentric coordinate system makes it possible to depict the originally 4D data (two-class confusion matrices) as points in 3D. In the following section, we will use this property to visualize measure gradients for different class proportions.

B. Class Proportion Gradient Components

Visualizations in the barycentric coordinate system have been already used to analyze entire ranges of classification measures by color-coding measure values [10]. Here, instead of static measure values we investigate the measures' dynamics by studying their gradients.

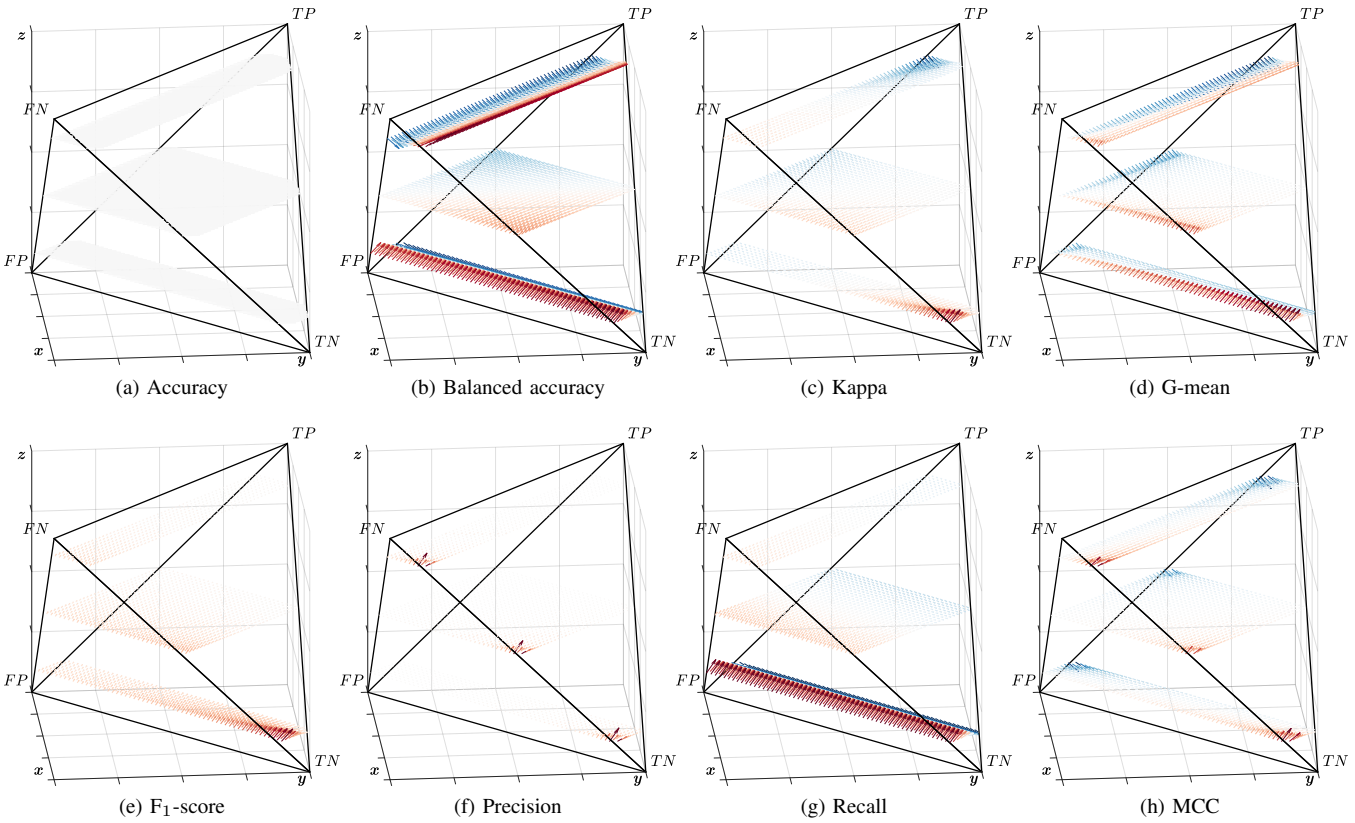
Since every possible confusion matrix and its corresponding measure *value* can be visualized as a *point* in the barycentric space, one can also calculate the *gradient* of the analyzed measure and depict it as a *vector*. The gradient shows the direction of the greatest rate of increase of the measure and its magnitude is the rate itself. In our case, this can be translated to the direction of changes in the confusion matrix that causes the greatest increase in the measure's value. To decipher the gradients of measures in the barycentric space, we first explain the meaning of their components.

Looking at Fig. 2, one can notice that the tetrahedron with vertices representing vectors TP, FP, FN, TN is placed in a 3D space defined by axes x, y, and z. Moving along the x-axis corresponds to moving from confusion matrices with all examples in vertices TP and FP to confusion matrices where all predictions are FN or TN. Recalling the notation from Table I, this means that the direction of the x-axis relates to changes in the proportions of classifier predictions: $\widehat{P} \xrightarrow{x} \widehat{N}$. Analogously, moving along the y-axis corresponds to changing the proportion of correct predictions in the confusion matrix ($F \xrightarrow{y} T$), whereas the z-axis can be associated with changes in class proportions ($N \xrightarrow{z} P$). Therefore, each z cross-section (Fig. 2) corresponds to all possible confusion matrices for a given imbalance ratio ir for a dataset of size n, whereas the entire tetrahedron encapsulates confusion matrices with n examples for all class ratios.

Another interpretation of the described directions x, y, z corresponds to rearranging the confusion matrix without changing the number of examples n. For the x-axis, the rearrangement involves changing the predictions of the classifier — moving examples from the left to the right column of the confusion matrix (Table I). On the other hand, the y-axis corresponds to changing incorrect predictions into correct ones. However, from the point of this study we are particularly interested in the movement along the z-axis, which can be implemented by relabeling negative examples to positive ones. Therefore, moving up and down the tetrahedron generally corresponds to varying class proportions, but if z changes while x and y remain constant the movement corresponds to example relabeling (concept drift) that also results in a change of class proportions. It is also worth noting that some preprocessing methods relabel training examples to enhance classifier predictions on imbalanced data [38].

Figure 3 presents gradients of the eight considered measures. To facilitate the analysis of z components of the gradients, we color-coded the gradients: red arrows have their z component pointing toward proportions with more positive examples (up), whereas blue arrows point toward proportions with more negative examples (down).

As the visualizations presented in Fig. 3 show, there are notable differences in the measure's gradients. Interestingly, *accuracy* is the only measure for which all the z gradient components are zero. This means that when the test data are relabeled in such a way that only the class proportion changes, the measure's value remains the same. *Precision* and F_1*-score*,



(a) Accuracy (b) Balanced accuracy (c) Kappa (d) G-mean

(e) F_1-score (f) Precision (g) Recall (h) MCC

Fig. 3: Measure gradients depicted in barycentric space. Gradients are represented as arrows with their length corresponding to the gradients norm and their color depicting the magnitude and direction of their 'vertical' element; red arrows have their vertical element pointing toward proportions with more positive examples (up), whereas blue arrows point toward proportions with more negative examples (down).

on the other hand, only have gradients with z components pointing towards distributions with more positive examples. Therefore, regardless of the class ratio, it is always, albeit not equally, beneficial in terms the values of these measures to relabel the test data in such a way that there are more positive examples. Since *recall* focuses on the positive class, for ratios with very few positives it is much more beneficial if the proportions change. On the other hand, the gradients of *balanced accuracy* indicate that for imbalanced data a change in class proportions is helpful when the classifier predicts mostly one class, ignoring the other. The remaining three measures (*G-mean*, *Kappa*, *MCC*) have symmetrical gradients with large z components mostly for confusion matrices where only the majority class is recognized correctly. In such cases, a change in class proportion will improve these measures' values.

To validate the presented findings, we performed an experiment involving dynamic class ratio changes in a data stream. Our hypothesis was that measure values obtained by an online classifier learning from a dataset with a dynamically changing imbalance ratio should reflect the presented gradients. To verify this hypothesis, we prequentially evaluated [23] a Hoeffding Tree classifier on two data streams created using the SEA generator in MOA [34]. The first stream (Fig. 4) contains two sudden class ratio changes (1:15/1:4/1:1) appearing after 30 k and 40 k examples; upper axis of the plot presents the ranges of each class proportion. The second stream (Fig. 5) was created analogously, but with negatives serving as the minority class (15:1/4:1/1:1).

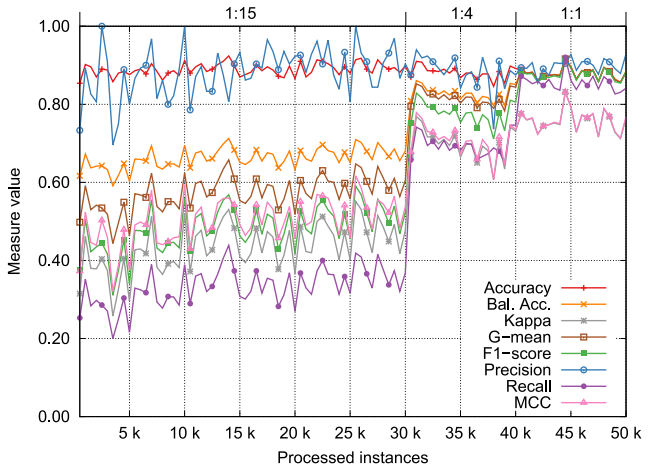


Fig. 4: The impact of changing the class ratio from 1:15 (underrepresented positives), through 1:4, to 1:1 (class balance). Results of prequential evaluation of a Hoeffding Tree classifier on a stream created using the SEA generator.

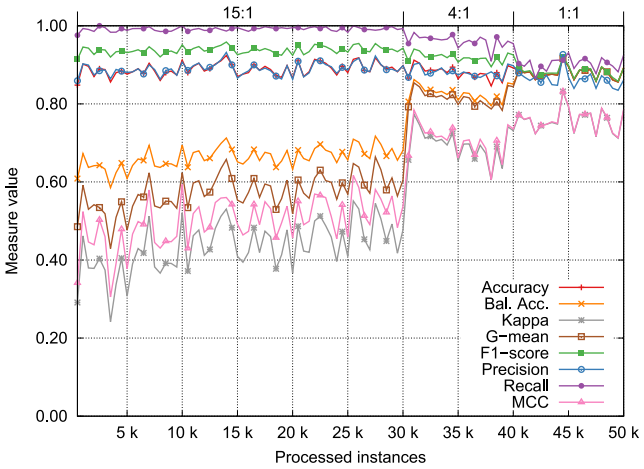


Fig. 5: The impact of changing the class ratio from 15:1 (underrepresented negatives), through 4:1, to 1:1 (class balance). Results of prequential evaluation of a Hoeffding Tree classifier on a stream created using the SEA generator.

Figures 4 and 5 confirm that the effect of class ratio changes is related to the measure's gradients. On both plots, *precision* and *accuracy* maintain almost the same level, regardless of the imbalance ratio. This directly translates to the gradients of these measures (Fig. 3), which have zero or close-to-zero z components. It is also worth noting that the variability of *precision* is in accordance with observations made on the measure's histograms — for ratios with a small number of positives high values of precision are underrepresented, thus, increasing the fluctuations of its values. F_1*-score* and *recall* have asymmetric gradients. For ratios with few positive examples the gradients have strong z components, whereas for datasets with many positive examples, the z component is close to zero. Due to this asymmetry, the changes in values of F_1*-score* and *recall* are much more prominent in Fig. 4 than in Fig. 5. Finally, the remaining measures (*balanced accuracy*, *Kappa*, *G-mean*, *MCC*) have symmetric gradients, therefore, their plots in Fig. 4 and Fig. 5 are exactly the same.

The experiments have shown a direct connection between measure gradients and the impact of sudden class ratio changes. We believe that the presented gradients could be also potentially used to estimate the effects of over- and under-sampling methods for classifiers producing concrete confusion matrices. By artificially modifying class proportions, sampling methods change the measures' dynamics. This connection, however, is still to be investigated and is out of the scope of this study.

V. Discussion

As related studies show, each of the eight analyzed measures represents different aspects of classification performance often leading to quite different interpretations [21]. This shows that there is no single measure that is the best choice in all situations. Nevertheless, the preformed study of measure dynamics provides some guidelines how to choose and use the analyzed measures in particular sub-cases.

One interesting observation from this study is that *recall* has a pmf that resembles a uniform distribution. Therefore, *recall* does not need any histogram-based normalization and can be safely compared between multiple static datasets. However, for very high imbalance ratios *recall* has a very small range of different values it can achieve. This phenomenon might be considered a problem when *recall* is used to for model selection or drift detection. Nevertheless, for data streams were new examples arrive continuously this issue is less important.

Precision and F_1*-score* are clearly two measures that require histogram-based normalization if their values are to be compared across multiple datasets. As experiments in Section III-B have shown, the interpretation of dataset difficulty can change dramatically after normalization. Moreover, *precision* and F_1*-score* only have gradients pointing towards distributions with more positive examples. Therefore, their response to dynamically changing class ratios is asymmetric and they are potentially more susceptible to simple oversampling than other measures. It is worth noting that the similarity of these two measures results from the fact that *precision* is "incorporated" in the definition of F_1*-score*.

Kappa and *MCC* behave very similarly for value ranges equal or greater than zero. Indeed, the histograms of these two measures have similar shapes for values between 0 and 1, and the gradients of both measures are also very similar. That is why, in experiments from Section III-B the effect of normalization was very similar for both measures. We note however, that this would not be true if the classifiers achieved very poor performance (below zero). In such cases, *MCC* seems to be a more symmetrical and more interpretable measure.

Finally, *accuracy*, *balanced accuracy*, and *G-mean* share some similarity in their $pmfs$, with most confusion matrices being associated with middle-range values. However, *balanced accuracy* and *G-mean* have stable histograms across varying class ratios, making these measures more suitable for tracking the performance of online classifiers trained on drifting imbalanced streams. Nevertheless, *accuracy* has a unique asset among all the studied measures — its gradients are independent of the class ratio. Therefore, if the fraction of correct predictions does not change, *accuracy* will not change. This is in contrast to other measures, which tend to react strongly to class ratio changes, as manifested by experiments in Section IV-B.

VI. Conclusions

The analyses presented in this paper have shown that changes in class proportions have direct effects on measure values. Moreover, we have demonstrated that measure value interpretation should not disregard the imbalance ratio, especially in data streams where the data is prone to concept drift or sudden class ratio changes. As the performed experiments have shown, these observed changes in measure behavior directly translate to the performance of drift detectors monitoring imbalanced streams. Consequently, our findings support the process of measure selection for a particular classification task and imbalance ratio. They help to interpret the measure values through the normalization method and raise the

awareness of possible value changes caused solely by evolving class proportions. In this context, the realization that values of *accuracy* and *precision* remain unchanged through class ratio changes, F_1*-score* and *recall* are susceptible to changes mainly when positives are underrepresented, and that *balanced accuracy*, *Kappa*, *G-mean*, *MCC* have symmetric gradients, can be essential for proposing new classifiers for imbalanced streams.

As future work, we plan to investigate the relation between measure gradients and specialized preprocessing methods for imbalanced data. Such preprocessing methods could take into account the class proportions, confusion matrix, and target evaluation measure. Moreover, it would be interesting to expand the analysis of measure dynamics to other classification scenarios, such as multi-label data, where minority and majority classes can concur [14], or window-based object detection using deformable part-base models, where the class proportions can vary significantly [12]. Finally, an interesting avenue of future research involves modeling the measure's probability mass function through general formulas, rather than by generating confusion matrices and counting value frequencies.

Acknowledgment

The authors would like to thank Maciej Piernik for insightful comments on a draft of this paper. This work was supported by PUT Institute of Computing Science Statutory Funds.

References

[1] P. Branco, L. Torgo, and R. Ribeiro, "A survey of predictive modeling under imbalanced distributions," *ACM Comput Surv*, vol. 49, no. 2, p. 31, 2016.

[2] H. He and Y. Ma, Eds., *Imbalanced Learning: Foundations, Algorithms and Applications*. IEEE - Wiley, 2013.

[3] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in AI*, vol. 5, no. 4, pp. 221–232, 2016.

[4] K. Napierala and J. Stefanowski, "Types of minority class examples and their influence on learning classifiers from imbalanced data," *J. Intell. Inf. Syst.*, vol. 46, no. 3, pp. 563–597, 2016.

[5] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

[6] B. Hu and W. Dong, "A study on cost behaviors of binary classification measures in class-imbalanced problems," *CoRR abs/1403.7100*, 2014.

[7] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, 2009.

[8] Q. Gu, L. Zhu, and C. Z., "Evaluation measures of the classification performance of imbalanced data set," in *Proc. ISICA*. Springer, 2009, pp. 461–471.

[9] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: An overview," *Bioinformatics*, vol. 16, pp. 412–424, 2000.

[10] D. Brzezinski, J. Stefanowski, R. Susmaga, and I. Szczęch, "Visual-based analysis of classification measures and their properties for class imbalanced problems," *Information Sciences*, vol. 462, pp. 242–261, 2018.

[11] A. Fernández, V. López, M. Galar, M. J. del Jesús, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowl.-Based Syst.*, vol. 42, pp. 97–110, 2013.

[12] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[13] F. Charte, A. J. Rivera, M. J. del Jesús, and F. Herrera, "Addressing imbalance in multilabel classification: Measures and random resampling algorithms," *Neurocomputing*, vol. 163, pp. 3–16, 2015.

[14] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Dealing with difficult minority labels in imbalanced mutilabel data sets," *Neurocomputing*, vol. 326-327, pp. 39–53, 2019.

[15] J. Stefanowski and D. Brzezinski, *Encyclopedia of Machine Learning and Data Mining*. Springer, 2017, ch. Stream Classification, pp. 1191–1199.

[16] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Wozniak, "Ensemble learning for data stream analysis: A survey," *Information Fusion*, vol. 37, pp. 132–156, 2017.

[17] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, pp. 44:1–44:37, 2014.

[18] S. Wang, L. L. Minku, and X. Yao, "Resampling-based ensemble methods for online class imbalance learning," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1356–1368, 2015.

[19] C. Drummond and R. C. Holte, "Cost curves: An improved method for visualizing classifier performance," *Machine Learning*, vol. 65, no. 1, pp. 95–130, 2006.

[20] R. Susmaga and I. Szczęch, "Can interestingness measures be usefully visualized?" *Int. J. Applied Math. Comp. Science*, vol. 25, no. 2, pp. 323–336, 2015.

[21] N. Japkowicz, *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, 2013, ch. Assessment Metrics for Imbalanced Learning, pp. 187–206.

[22] M. Bekkar, H. Djemaa, and A. Taklit, "Evaluation measures for models assessment over imbalanced data sets," *Journal of Inform. Eng. and Appl.*, vol. 3, no. 10, pp. 27–38, 2013.

[23] J. Gama, R. Sebastião, and P. P. Rodrigues, "On evaluating stream learning algorithms," *Mach. Learn.*, vol. 90, no. 3, pp. 317–346, 2013.

[24] D. Brzezinski and J. Stefanowski, "Prequential AUC: Properties of the area under the roc curve for data streams with concept drift," *Knowledge and Information Systems*, vol. 52, no. 2, pp. 531–562, 2017.

[25] D. M. Powers, "What the F-measure doesn't measure: Features, flaws, fallacies and fixes," *CoRR abs/1503.06410*, 2015.

[26] P. A. Flach, "The geometry of ROC space: Understanding machine learning metrics through ROC isometrics," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 194–201.

[27] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[28] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 233–240.

[29] R. Caruana and A. Niculescu-Mizil, "Data mining in metric space: An empirical analysis of supervised learning performance criteria," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 69–78.

[30] R. Alaíz-Rodríguez, N. Japkowicz, and P. E. Tischer, "A visualization-based exploratory technique for classifier comparison with respect to multiple metrics and multiple domains," in *Proc. 19th European Conf. Mach. Learn., Part II*, 2008, pp. 660–665.

[31] J. Hernández-Orallo, P. A. Flach, and C. F. Ramirez, "Brier curves: a new cost-based visualisation of classifier performance," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 585–592.

[32] T. T. Soong, *Fundamentals of Probability and Statistics for Engineers*. Wiley, 2004.

[33] G. Weiss, *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, 2013, ch. Foundations of Imbalanced Learning, pp. 13–43.

[34] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: Massive Online Analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 1601–1604, 2010.

[35] D. Brzezinski and J. Stefanowski, "Prequential AUC for classifier evaluation and drift detection in evolving data streams," in *New Frontiers in Mining Complex Patterns*, ser. LNCS, vol. 8983, 2015, pp. 87–101.

[36] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[38] K. Napierala, J. Stefanowski, and S. Wilk, "Learning from imbalanced data in presence of noisy and borderline examples," in *Proc. 7th Int. Conf. Rough Sets Current Trends Comput.*, 2010, pp. 158–167.

Dariusz Brzezinski received his M.Sc. and Ph.D. degrees in Computer Science from Poznan University of Technology, Poland, in 2010 and 2015 respectively. He is currently an assistant professor in the Department of Data Processing Technologies at Poznan University of Technology. His research interests include data stream mining, concept drift, evaluation measures, and machine learning applications in structural crystallography.

Jerzy Stefanowski is an Associate Professor in the Institute of Computing Science, Poznan University of Technology, Poland. He received the Ph.D. and Habilitation degrees in computer science from this university. His research interests include machine learning, data mining and intelligent decision support—in particular rule induction, multiple classifiers, class imbalance, data preprocessing, and data streams.

Robert Susmaga received his M.Sc., Ph.D. and D.Sc. degrees in computing science from the Poznan University of Technology, Poland, in 1994, 2001 and 2015, respectively, where he has been working at the Laboratory of Intelligent Decision Support Systems, Institute of Computing Science, ever since. His research interests focus on various elements of machine learning, knowledge discovery and data mining, and include the analysis and visualization of multidimensional data.

Izabela Szczech received her B.Eng., M.Sc. and Ph.D. degrees in computing science from the Poznan University of Technology, Poland in 2002, 2004 and 2008, respectively. Currently she is an assistant professor at the same university, at the Laboratory of Intelligent Decision Support Systems, Institute of Computing Science. She works mainly on topics related to data mining, in particular to measures of rule interestingness and their properties.