F_β-plot - a visual tool for evaluating imbalanced data classifiers

Szymon Wojciechowski[*] szymon.wojciechowski@pwr.edu.pl
Michal Wozniak michal.wozniak@pwr.edu.pl
Wroclaw University of Science and Technology, Wybrzeze Wyspianskiego 27, 50-370,Wroclaw, Poland

Editors: Manuel Grana, Pawel Ksieniewicz, Leandro Minku, Pawel Zyblewski

Abstract

Imbalanced data classification suffers from a lack of reliable metrics. This runs primarily from the fact that for most real-life (and commonly used benchmark) problems, we do not have information from the user on the actual form of the loss function that should be minimized. Although it is pretty common to have metrics indicating the classification quality within each class, for the end user, the analysis of several such metrics is then required, which in practice causes difficulty in interpreting the usefulness of a given classifier. Hence, many aggregate metrics have been proposed or adopted for the imbalanced data classification problem, but there is still no consensus on which should be used. An additional disadvantage is their ambiguity and systematic bias toward one class. Moreover, their use in analyzing experimental results in recognition of those classification models that perform well for the chosen aggregated metrics is burdened with the abovementioned drawbacks. Hence, the paper proposes a simple approach to analyzing the popular parametric metric F_β. We point out that it is possible to indicate for a given pool of analyzed classifiers when a given model should be preferred depending on user requirements.

Keywords: Imbalanced data, Real-world applications, Automated machine learning

1. Introduction

The problem of classifier evaluation has been known and addressed in the literature for years (Japkowicz and Shah, 2011). One of the critical issues is the selection of appropriate quality metrics and the proposed experimental protocol. This work will address the first problem for the imbalanced data classification. In this task (we will consider the most popular binary problem), we are dealing with the difference in abundance between classes, specifically referring to the dominant class as the majority class and the less abundant class as the minority class. Furthermore, we typically presume that an error made in the minority class has a greater cost than an error made in the majority class. The difficulty in classifying imbalanced data arises primarily as we lack data on the costs of these errors, known as the loss function.

Usually, two simple metrics that indicate errors made on a given class are analyzed, i.e., *specificity* or *sensitivity* (also called *recall*), or *recall* and *precision*, which indicates how many objects are classified in a minority class. Of course, to say anything about the quality of a given classifier, we need to analyze at least two of these metrics - *recall* and *precision*, which are usually used nowadays. Unfortunately, we can't determine the best

[*] Corresponding author.

classifier without details about the loss function or the significance of each metric in the user's context. Additionally, following the common human inclination to communicate information with a single numerical indicator, many aggregated metrics have been suggested, merging information about *precision* and *recall*, or *specificity* and *sensitivity* into an aggregated metric. Unfortunately, most people do not realize that such an approach has serious drawbacks, i.e., such metrics are ambiguous (they do not indicate which values of the simple metrics characterize the evaluated model), and also choosing them as a criterion often leads to the selection of classifiers biased towards the majority class.

Most researchers employ computer experiments to evaluate the predictive performance of proposed models. Usually, the results are promising, and the authors conclude that their classifiers can outperform the state-of-the-art algorithms. However, it should be noted that Wolpert's *no free lunch* theorem concludes that the best learning algorithm does not exist (Wolpert, 2001), so any conclusions drawn from experiments are conditioned on given datasets and a chosen experimental protocol, which specifies, among other things, which metrics are taken into account during the comparison.

Moreover, imbalanced data classification has become an arena for fighting for the best values of selected metrics, which, as we pointed out, can be biased. According to Goodhart's law (Strathern, 1997), "When a measure becomes a target, it ceases to be a good measure". Unfortunately, finding the right metrics for classifying imbalanced data is very difficult because we generally do not know how costly the errors made on the different classes are. Often, the assumptions made in the experiments about these costs (expressed in the assumed set of metrics) are far from reality, and the authors of the papers do not bother to show comparisons for other cases, i.e., for different costs of the mentioned errors.

Such an approach could be detrimental to the development of classification methods, as only solutions for a specific relatively standard set of metrics, which express particular (not always true) expectations about the quality of classifiers, are preferred and published. This results in solutions that do not perform well for a given set of metrics not being published and thus not being promoted even if they could be helpful in other metrics settings and thus find application in specific practical problems.

Of course, one should not be surprised by this approach, as the primary motivation of most researchers is to publish their work. However, most journals are not interested in negative results because they are not as attractive to readers and do not seem interesting [1]. This means that even if we design a classifier evaluation experiment according to the guidelines for classifier testing such as (Demšar, 2006; García and Herrera, 2009; García et al., 2010), there is still room for metric manipulation (Stapor et al., 2021).

This work focuses on a proposal to evaluate the classifier quality using the parametric measure F_β, which is commonly chosen for the value $\beta = 1$ and does not follow the end-user expectations. The criticism of $F1$ or $F - score$ can be addressed since the parameter β *de facto* indicates how much more critical *recall* is compared to *precision* (Hand and Christen, 2018).

This work contributes to imbalanced data classifier evaluation, especially it points out the problem of inadequate comparison of classification methods, which, through an improper choice of metrics, promotes models with characteristics suitable for only particular user

[1] Fortunately, more and more people see the value in publishing negative results as well, and an increasing number of journals are choosing to do so (see, e.g., https://doi.org/10.15252/embr.201949775)

Table 1: Confusion matrix for a two-class problem.

		PREDICTED CLASS	
		POSITIVE	NEGATIVE
TRUE CLASS	POSITIVE	*True Positive* (TP)	*False Negative* (FN)
	NEGATIVE	*False Positive* (FP)	*True Negative* (TN)

preferences, i.e., a specific preference for the cost of errors made on different fractions of the data. We propose a simple visualization tool indicating which models are helpful depending on the user's expectations. The usefulness of such an analysis is demonstrated using a selected benchmark analysis as an example.

2. Motivations

This section focuses on selected properties of the metrics while realizing that this section aims not to provide an exhaustive overview of the metrics but to indicate the motivation for developing a method that allows fair comparison of imbalanced data classifiers.

Many metrics for classifier evaluation have been proposed (Japkowicz and Shah, 2011). Usually, they are calculated based on *confusion matrix*, which summarizes the number of instances from each class classified correctly or incorrectly as the remaining classes. For a two-class classification task, consider the 2 × 2 confusion matrix (see Tab. 1).

The most popular metric is the *Accuracy* (*Acc*):

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

However, it is easy to show that *Acc* is heavily biased towards the majority class and could lead to misleading conclusions. It is a good metric only for tasks that use the so-called $0 - 1$ loss function, i.e., when the cost of errors committed on each class is the same. This observation has led the imbalanced data science community to look for other metrics.

Often we are interested in classifier evaluation on only a part of the data, i.e, positive or negative data. *True Positive Rate* (*TPR*) also known as *Recall* or *Sensitivity*), *True Negative Rate* (*TNR*) known as *Specificity*, and *Positive Predictive Value* (PPV) also called *Precision*:

$$TPR = \frac{TP}{TP + FN}, \tag{2}$$

$$TNR = \frac{TN}{TN + FP}, \tag{3}$$

$$PPV = \frac{TP}{TP + FP} \tag{4}$$

It is well known that people do not like to make comparisons based on multiple criteria (metrics). Moreover, solving such a problem leads to multi-criteria optimization, which may return not one but several solutions, so-called non-dominated solutions, i.e., solutions for which none of the metrics can be improved without degrading some of the other criteria. Usually, to select a solution that suits the user's preferences, a multiple criteria decision analysis (MCDA) is employed (Cinelli et al., 2020). However, that is a difficult task that has been researched for years. MCDA solutions focus on designing a decision-making process to assist the user in deciding. Such a decision-making process primarily helps identify the user's preferences, which could be used in a decision model. An example of such a process is *PROMETHEE*, which relies on pairwise comparisons to rank alternatives evaluated on multiple criteria (Mareschal and Brans, 2005).

An alternative approach is to reduce all criteria to a single one, a function of all used criteria, and select a solution according to its value. This approach has been widely accepted in the scientific community and focuses on imbalanced data classification (Luque et al., 2019).

Many metrics have been proposed, generally aggregating *TPR* and *TNR*, or *TPR* and *PPV*. Such metrics include arithmetic, geometric, or harmonic means between the two components: *recall* and *specificity*. There are also other proposals trying to enhance one of the two components of the mean, for example, *Index of Balanced Accuracy* (García et al., 2009) or F_β*score* (Sokolova and Lapalme, 2009). This work focuses on using F_β, thus let us present its definition

$$F_\beta = \frac{(\beta^2 + 1) \times PPV \times TPR}{\beta^2 \times PPV + TPR} \tag{5}$$

The β expresses the trade-off between selected simple metrics, i.e., how much more critical TPR is to the user than PPV. Improper selection of the value of the mentioned parameter can lead to the choice of an inappropriate classifier, e.g., favor the majority class for an imbalanced data classification task (Brzezinski et al., 2018).

Interestingly, many practical recommendations for quality metrics note the importance of the β parameter and suggest using indicators for several values (usually 0.5, 1, and 2). Unfortunately, papers comparing the quality of classifiers of imbalanced data provide only the F_1 typically.

3. F_β-plot analysis

The idea of F_β-plot is to visualize the F_β values for each of the analyzed classifiers depending on the β value and then determine the β ranges to indicate which classifier takes the best values.

The ranking of the evaluated methods will vary with the value of the β parameter. Classifiers with a preference for the majority class (and, therefore, achieving a higher precision value) will be superior as the value of the β parameter decreases. While the β is converging to 0, the F_β value will approach the *precision* score of the classifier. Similarly, the F_β value will approximate the TPR component, with a higher value of the β parameter increasing to infinity. The point of balance between the components is 1 – the value for which the F_β function is equivalent to the harmonic mean.

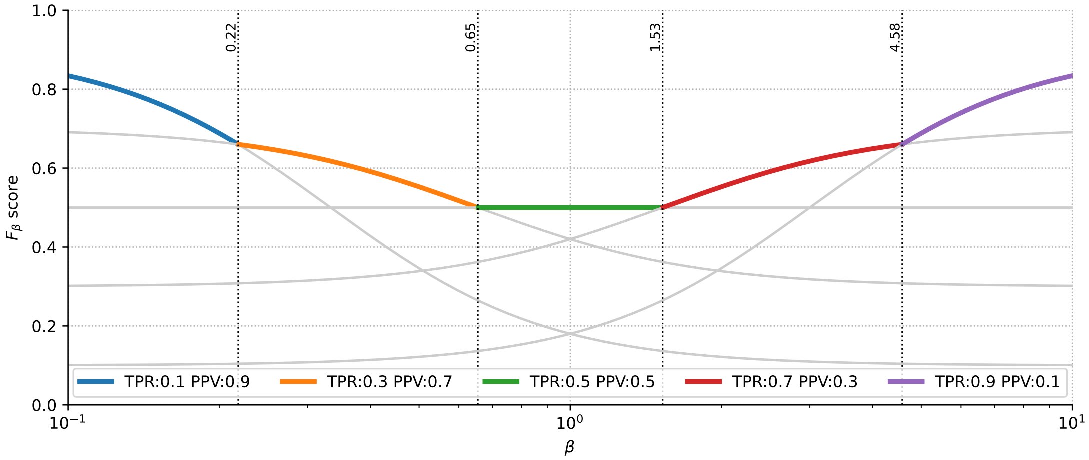

Figure 1: Example of relations between F_β and β

F_β-plot is a tool for observing changes in ranking depending on the β parameter. Figure 1 shows a simulation of such a relationship. The β values are presented on a logarithmic scale to make the results more readable.

Five different scenarios considering different configurations of TPR and PPV were simulated. As one can observe, it is affirmed that for values of $\beta = 1$, a perfectly balanced classifier is preferred. However, even with slight differences in the β parameter's value, the curves' ranking changes significantly, preferring results more biased towards one of the parameters. Biased solutions are preferred as well for strongly deviating β values.

Noteworthy is the fact that β values where curves are intersecting are easily determined

$$\frac{(1+\beta^2)\cdot \mathrm{PPV}_A \cdot \mathrm{TPR}_A}{(\beta^2 \cdot \mathrm{PPV}_A) + \mathrm{TPR}_A} = \frac{(1+\beta^2)\cdot \mathrm{PPV}_B \cdot \mathrm{TPR}_B}{(\beta^2 \cdot \mathrm{PPV}_B) + \mathrm{TPR}_B} \tag{6}$$

$$\beta = \sqrt{\frac{(\mathrm{TPR}_A \cdot \mathrm{TPR}_B \cdot (\mathrm{PPV}_B - \mathrm{PPV}_A)}{(\mathrm{PPV}_A \cdot \mathrm{PPV}_B \cdot (\mathrm{TPR}_A - \mathrm{TPR}_B)}} \tag{7}$$

where indexes A and B refer to exemplary classifiers.

Nevertheless, the representation of the achieved scores in their entire range, along with an indication of the intersections (thus, ranking changes), allows a more thorough evaluation of the quality of tested models. At the same time, the F_β-plot provides guidelines to the system designer – indicating the method that achieves the best performance according to the preference of the system's end user, which should determine the proportion between the cost of errors.

Interpreting F_β-plot

To better explain and give practical example of F_β-plot, we conducted simple experiment considering a well-known imbalanced dataset – *Thyroid Disease* (Quinlan, 1987). We trained

92 models using k-nearest neighbors algorithm (kNN) combined with various SMOTE-based oversamplers presented in (Kovács, 2019).

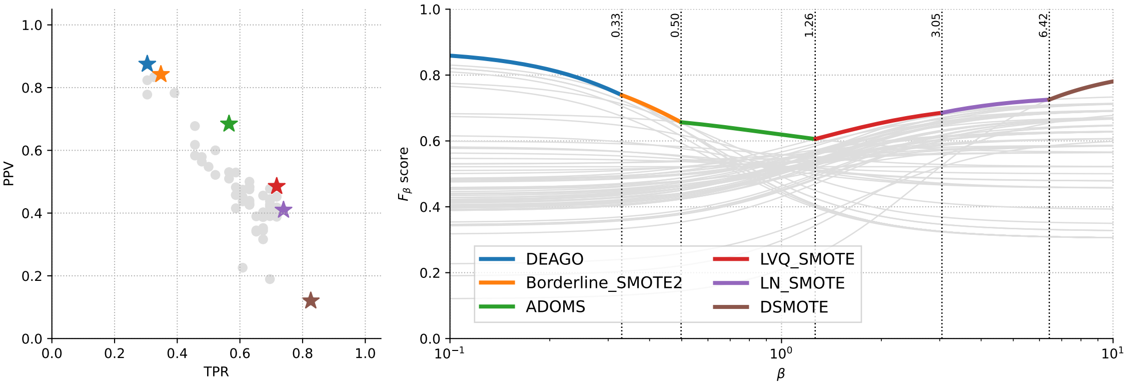

Figure 2: F_β-plot for *Thyroid Disease* with hold-out evaluation.

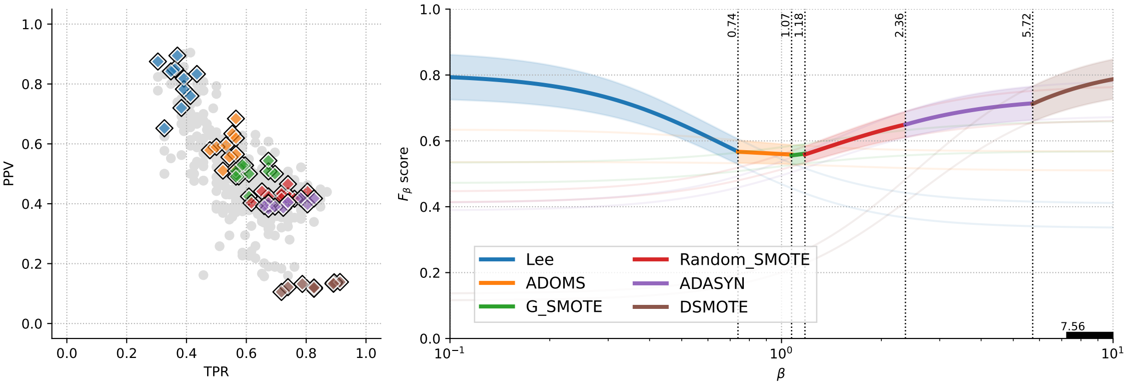

Figure 3: F_β-plot for *Thyroid Disease* with cross-validation.

Firstly, the performance of the models was determined on a 20% hold-out choosen at random from dataset. The F_β-plot for the obtained results is presented in Figure 2, along with a scatter plot of TPR and PPV of selected methods. The colored elements indicate the methods that achieved the best result for changing β value in analysis range.

We may observe that, depending on the configuration of the β parameter, it is possible to distinguish six models for which result in the best in terms of F_β value. The most balanced method ($\beta = 1$) is the `ADOMS` oversampler, for which we can also notice, that it has a slight majority class preference. The other discovered algorithms that prefer the majority class are `Borderline_SMOTE2` and `DEAGO`, although their curve characteristics are similar and are reflected in the close distance in the scatter plot. From $\beta = 1.26$, which we can describe as a preference towards the minority class, the `LVQ_SMOTE`, `LN_SMOTE`, and `DSOMTE` are the best performing algorithms. We should also point out that the selected models constitute a boundary in TPR - PPV space and it is relative to the point $(1, 1)$ - which would be considered a *ideal* model.

Statistical analysis on F_β-plot

Many experiments use more rigorous experimental protocols commonly based on k-fold cross-validation. F_β-plot can be easily adapted to such protocol by extending plots of standard deviation values and proper region marking. An example is presented in Figure 3. The plot now includes standard deviations, and to explain their meaning, the scatter plot marked all values obtained by the considered method. The colored plots indicate the methods whose average F_β was the highest in the given range, and the black strip on the bottom axis marks a range in which the selected method is the best and the results are statistically significant (considering *t-test* with the $\alpha < 0.05$).

As expected, the list of selected methods has changed over repeated evaluations. The reoccurring algorithms are `DSMOTE` and `ADOMS`. However, for other oversamplers, the ranking has changed. Better or similar ones replaced the previously selected methods. The observed change reflects using a broad set of oversampling algorithms, which tend to follow similar principles, thus generating similar synthetic samples for the model to be prepared. To distinguish statistically better algorithms on the sampled interval, following the results of the performed statistical test is essential. It can be determined that in the case of an extreme preference against recall ($\beta > 7.56$), `DSMOTE` based models are the only ones to achieve the TPR oriented result. This information might be practical and useful if the system designer aims to provide a system biased toward the minority classes (e.g., a screening test system). As for the other preferences, we have to consider that there is a model that might act similar to the one marked on F_β-plot. Nevertheless, based on the statistical tests, it is also possible to indicate a subset of similar methods. Consequently, for a practical example, it is also possible to consider their computational complexity to select the best-suited method.

4. Experiments

The experimental study will present the characteristics of F_β plots for selected benchmark imbalanced datasets (Derrac et al., 2015). A vast pool of oversampling algorithms will be compared (Kovács, 2019), which will be used to preprocess the data, followed by the k-NN classifier. For the evaluation protocol, we choose *2x5-fold stratified cross-validation*. Table 2 presents the eight datasets chosen for this experiment. The code for experiments reproduction, as well as F_β plot code, is publicly available [2]. The experiment aims to investigate observable relationships and discuss their interpretation. The experiment results are presented in Figure 4.

For most datasets, no statistically significant best classifier was observed throughout the analyzed range, and it is only possible to highlight `AMSCO` on the $\beta > 2.36$ interval for the *vehicle1* dataset. Other methods achieve the best values on the analyzed ranges, but it should be remembered that other oversampling can be identified to obtain similar results.

Additionally, it could be observed that for most problems, the value of the F_β remains high, not falling below 0.6. However, the exception is the strongly imbalanced set of *poker-8-*

[2] https://github.com/w4k2/fb-plot

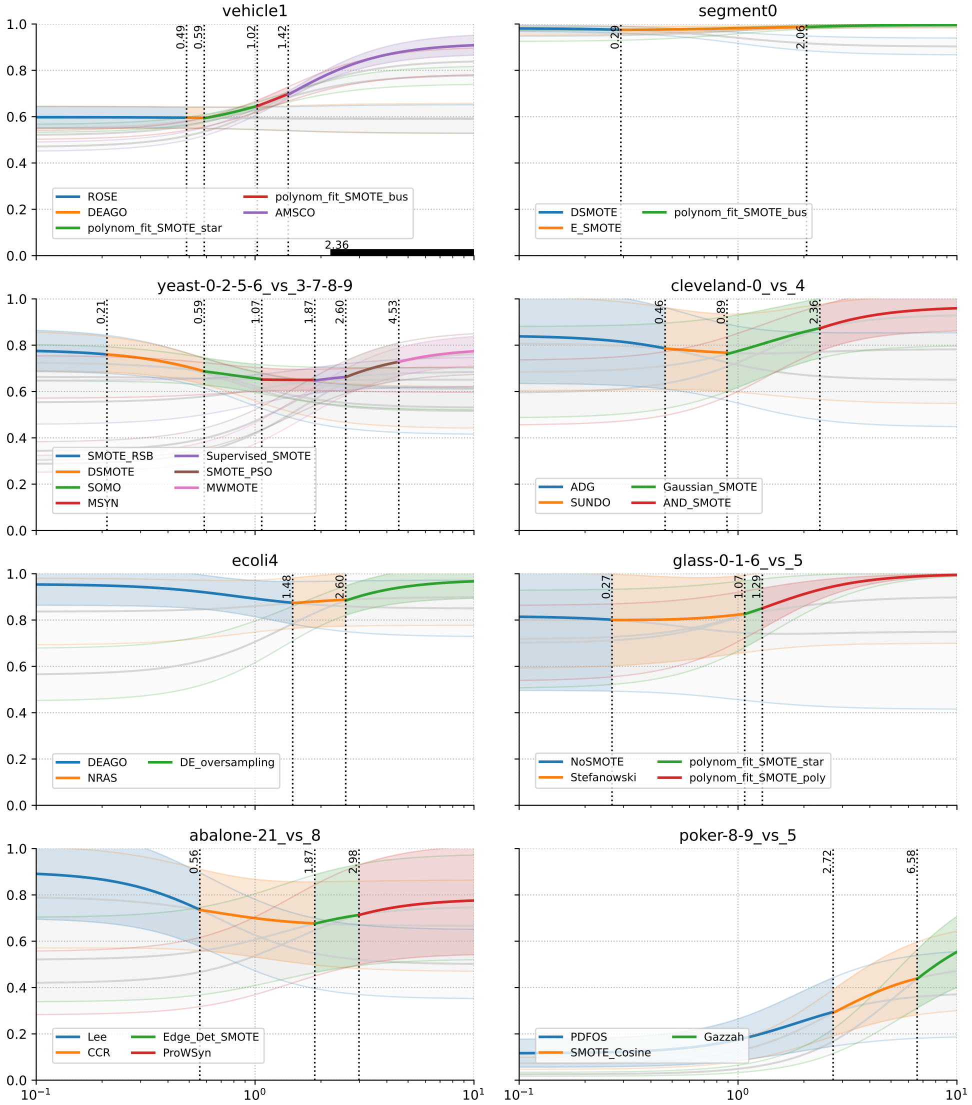

Figure 4: F_β-plots of selected datasets

9_vs_5, for which it is difficult to identify a method to achieve a satisfactory result, preferring precision. Such behavior may be related to the pool of preprocessing methods themselves, which - by design - seek to equalize (or outweigh) the model's bias against the minority class. Another factor that may also affect the observed result is the problem's difficulty, which is not always related to the degree of imbalance, but, for example, the number of objects of the borderline class (Napierala and Stefanowski, 2012; Skryjomski and Krawczyk, 2017). A similar effect is also observed for the sets *vechicle1* and *glass-1-6_vs_5*, and the opposite effect is observed for the set *segment0*, where we can almost always indicate the method for which the F_β is near 1. As for the other sets, the curve assembled from best

Table 2: Main characteristics of the chosen benchmark datasets

Name	Samples	Features	IR
vehicle1	846	18	2.90
segment0	2308	19	6.06
yeast-0-2-5-6_vs_3-7-8-9	1004	8	9.14
cleveland-0_vs_4	173	13	12.31
ecoli4	336	7	15.80
glass-0-1-6_vs_5	184	9	19.44
abalone-21_vs_8	581	8	40.50
poker-8-9_vs_5	2075	10	82.00

algorithms forms a ”V” shaped curve, where the lowest F_β values are achieved in the near surroundings of $\beta = 1$ with a slightly higher tendency towards the minority class.

5. Conclusion

The issue of selecting suitable metrics for imbalanced data problems remains pertinent. On one hand, it is recommended to avoid aggregated metrics and instead rely on simple metrics for analysis. On the other hand, it should be acknowledged that analyzing multiple criteria simultaneously may be challenging, especially for less experienced users or those who cannot determine the costs of incorrect decisions related to selected data fractions. This paper presents a simple $F_\beta - plots$ method to visualize the results of the experiments, allowing simultaneous evaluation of the quality of multiple methods for different values of β, i.e., the end user's expectation of the validity of TPR against PPV. The $F_\beta - plots$ method identifies the costs at which a particular method is beneficial. Additionally, it indicates the tasks for which a given classifier may be suitable, such as values of imbalance ratio that are proportional to costs between simple metrics, as suggested by Brzezinski et al. (Brzezinski et al., 2020). Such an analysis provides a broader perspective on the quality and scope of the tested classifiers.

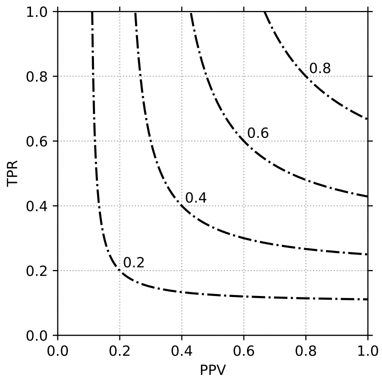

Figure 5: Example of different F_1 values in PPV-TPR space.

However, selecting an aggregate metric can be challenging due to its limited interpretability (Hand et al., 2021) and potential ambiguity, i.e., even if we use the different β

values, we still face the problem of aggregated metric ambiguousness because the same F_β value may be taken for different PPV and TPR values (see Figure 5). Thus, one might suspect that machine learning methods that use such an aggregated metric as a criterion will be biased toward specific values of simple metrics without providing the information that there are equally good solutions (in terms of a given metric) for other values of PPV and TPR. Additionally, F_β ignores the number of true negatives (Christen et al., 2023). The abovementioned issues were not discussed in this paper and are still waiting to be properly addressed.

Acknowledgments

This work was supported by the Polish National Science Centre under the grant No. 2019/35/B/ST6/04442 and the statutory funds of Department of Systems and Computer Networks of Wroclaw University of Science and Technology.

References

Dariusz Brzezinski, Jerzy Stefanowski, Robert Susmaga, and Izabela Szczch. Visual-based analysis of classification measures and their properties for class imbalanced problems. *Information Sciences*, 462:242 – 261, 2018. ISSN 0020-0255.

Dariusz Brzezinski, Jerzy Stefanowski, Robert Susmaga, and Izabela Szczech. On the dynamics of classification measures for imbalanced and streaming data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):2868–2878, 2020.

Peter Christen, David J. Hand, and Nishadi Kirielle. A review of the f-measure: Its history, properties, criticism, and alternatives. *ACM Comput. Surv.*, 56(3), oct 2023. ISSN 0360-0300.

Marco Cinelli, Miłosz Kadziński, Michael Gonzalez, and Roman Słowiński. How to support the application of multiple criteria decision analysis? let us start with a comprehensive taxonomy. *Omega*, 96:102261, 2020. ISSN 0305-0483.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

J Derrac, S Garcia, L Sanchez, and F Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *J. Mult. Valued Logic Soft Comput*, 17:255–287, 2015.

Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Inf. Sci.*, 180(10): 2044–2064, 2010.

Vicente García, Ramón Alberto Mollineda, and José Salvador Sánchez. Index of balanced accuracy: A performance measure for skewed class distributions. In *Iberian conference on pattern recognition and image analysis*, pages 441–448. Springer, 2009.

Salvador García and Francisco Herrera. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, 9:2677–2694, 2009. URL http://www.jmlr.org/papers/volume9/garcia08a/garcia08a.pdf.

David Hand and Peter Christen. A note on using the f-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28(3):539–547, may 2018. ISSN 0960-3174.

David Hand, Peter Christen, and Nishadi Kirielle. F*: an interpretable transformation of the f-measure. *Machine Learning*, 110, 03 2021.

Nathalie Japkowicz and Mohak Shah, editors. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011. ISBN 9780521196000.

György Kovács. smote-variants: a python implementation of 85 minority oversampling techniques. *Neurocomputing*, 366:352–354, 2019. (IF-2019=4.07).

György Kovács. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing*, 83:105662, 2019. ISSN 1568-4946.

Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, 2019. ISSN 0031-3203.

B Mareschal and JP Brans. Promethee methods. *International Series in Operations Research and Management Science*, 78:163–195, 2005.

K. Napierala and J. Stefanowski. Identification of different types of minority class examples in imbalanced data. In *Hybrid Artificial Intelligent Systems*, volume 7209 of *Lecture Notes in Computer Science*, pages 139–150. Springer Berlin Heidelberg, 2012.

Ross Quinlan. Thyroid Disease. UCI Machine Learning Repository, 1987.

Przemysław Skryjomski and Bartosz Krawczyk. Influence of minority class instance types on smote imbalanced data oversampling. In *first international workshop on learning with imbalanced domains: theory and applications*, pages 7–21. Pmlr, 2017.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

Katarzyna Stapor, Paweł Ksieniewicz, Salvador García, and Michał Woźniak. How to design the fair experimental classifier evaluation. *Applied Soft Computing*, 104:107219, 2021. ISSN 1568-4946.

Marilyn Strathern. 'improving ratings': audit in the british university system. *European Review*, 5(3):305–321, 1997.

David H. Wolpert. The supervised learning no-free-lunch theorems. In *In Proc. 6th Online World Conference on Soft Computing in Industrial Applications*, pages 25–42, 2001.