Probabilistic Metric to measure the imbalance in multi-class problems

Solander Patricio Lopes Agostinho SOLANDER.AGOSTINHO@UCAN.EDU
João Mendes-Moreira JMOREIRA@FE.UP.PT
LIAAD-INESC TEC, Universidade do Porto, R. Dr. Roberto Frias, 4200-465 Porto, Portugal
CATHOLIC UNIVERSITY OF ANGOLA, Av. Pedro de Castro Van-Dúnem Loy 24, Luanda

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

Abstract

In machine learning, imbalanced data has been one of the most relevant issue that the classifiers have to deal with. The most common techniques applied in this scenario are all, somehow, based on oversampling or under sampling concepts, In the former, the number of instances of minority classes are, somehow, increased while in the latter, the number of instances in the majority classes are somehow reduced. By increasing Pre-processing, approaches as the ones described have been well succeeded in binary classification problems.However, as the larger the number of classes, less effective the pre-processing approaches are. Another related problem is that the metrics that evaluate the predictive performance of the classifiers can be not effective in the presence of imbalanced data. The metrics used to measure the predictive performance of classifiers, can be divided into three groups: threshold, ranking and Probabilistic metrics. This paper aimed to purpose a probabilistic metric with the main objective of, given the results of a classifier in a multi-class domain, verify the relation between these result and the imbalance problem. The main purpose of this work, is to build a probabilistic metric based on non-parametric approaches, to measure the effect of imbalance feature of dataset in multi-class problems. As part of the work, a comparison with the existing metrics will be implemented and analyzed, both to understand the relation between them and to choose the best of them according to each scenario.

Keywords: imbalanced data, multi-class domain, classification, probabilistic metric

1. Introduction

Nowadays, the imbalanced problem occurs in many fields, such as finances, manufacture, health, social media and others. According to Branco et al. (2017), most of the proposal works with the main purpose of measuring the predictive performance results using a metric, and most of them are well applied on binary problems(Gu et al., 2009). The scenario is different with multi-class imbalance problems, where most of the existing metrics tend to fail or give a not real performance. This domain has receiving more attention in present times, once the challenge has grown in present times (Wardhani et al., 2019). In terms, the common methods for imbalanced domain, according to the most recent studies, are concentrated in two types. The first is focused on improving the classifiers, by trying to reduce the sensitivity of them when they work with imbalanced data. In general, ensemble learning (Tian and Wang, 2017) and cost sensitive algorithms (Ling and Sheng, 2010) are

approaches with the main concern to increase the robustness of the classifier or make the cost of misclassifying the minority class higher than the majority class Zheng et al. (2020).

In another hand, we have the approaches focused on sampling methods, with two basic ideas, oversampling and under sampling methods. Methods like Synthetic Minority Oversampling Technique (SMOTE)(Jeatrakul et al., 2010), Adaptive Synthetic Sampling (ADASYN), Random under-sampling (RUS), Condensed nearest Neighbor rule + Tomek links (CNNTL), Class Purity Maximization (CPM), Selective Preprocessing of Imbalanced Data (SPIDER) and Generative Adversarial Networks (GAN) (Kaur et al., 2017)(Wang et al., 2022) are examples that combine these methods and that are nowadays commonly used to deal with imbalance problems. Inside ensemble topics, two ways have been largely explored: bagging, also known as bootstrap aggregating, and boosting(Wardhani et al., 2019). The boosting approach has been largely used in present times, being one of the most successful ensemble approaches to deal with imbalanced data sets. The basic idea of boosting algorithms is to convert the weak learners to strong ones, and increase or improve their accuracy of prediction (Wang et al., 2006). The motivation scenario is, given a data set, after applying a classifier K_i and get the v_i accuracy, doing it with different classifiers and getting models with possible different values of accuracy, its mean, for $\{K_1, K_2, ..., K_i, ..., K_m\}$ where m is the number of classifiers applied resulting in $\{v_1, v_2, ..., v_i, ..., v_m\}$ values of accuracy, and assuming that each accuracy value is not equal than the rest, so we have possible different values of accuracy like $v_i, v_j \in V$, and $v_i \neq v_j$, and the idea is to use a method to combine all these models to get the final prediction (Deb et al., 2020).

1.1. Evaluation concerns

After knowing the ways to concern about the imbalanced problem in multi-class domains, the next step is to evaluate whether the classifier is giving the right results. Predictive performance evaluation metrics are used to evaluate the classification algorithms. According to Jason Brownlee (2020), for classification problems, metrics are used to compare the expected class label to the predicted class label or interpreting the predicted probabilities of the classes of the problem. About taxonomy, these metrics can be classified into three groups: Threshold, Ranking and Probabilistic metrics. The first group address to quantify the error generated by prediction tasks in classifiers. The common threshold metrics are accuracy given by $acc = \frac{[Correct\_predictions]}{[Total\_predictions]}$, and error metric, wrote by a complement of accuracy, given by $err = \frac{[Incorrect\_predictions]}{[Total\_predictions]}$. These kinds of metrics, despite being largely used, are inappropriate for imbalanced domains, once the high accuracy or low error might be due to the difficulty of the classifier to predict the minority classes, or predict more the majorities. Threshold metrics, more properly for the imbalanced case, are sensitivity-specificity and precision-recall, once they focus on each single class (Jason Brownlee, 2020). Outside the most popular threshold metrics, there are others approaches such as Kappa, Macro-Average Accuracy, Mean-class-Weighted Accuracy, Optimized precision and many others from old or recent studies. By another side, the ranking metrics addressed to evaluate classifiers according to how effective they can target at separating classes. These kinds of metrics are more useful in situations that the task is to select the best n instances of a dataset. These metrics are commonly based on Receiver Operating Characteristic (ROC) ROC AUC (Area under curve) (Prati et al., 2008).

This research work aims to propose a newly probabilistic metric, to measure the relation between the imbalanced problem and the performance of the classifiers. This approach has as main contributions: i) to determinate the effect of the imbalance problem in the results of the classification methods ii) to give inputs to evaluate the misclassification of the classifiers, iii)to discuss the way to choose the right evaluation metric according to the evaluation objective, iv) to provide a framework to apply the datasets and investigate the imbalanced effects.

2. Evaluation metrics for multi-class imbalanced domain

Evaluation metrics are essential to verify the trust about the classifier results. Accuracy, confusion matrix, precision, recall, F1 score, sensitivity, specificity, ROC curve and AUC are the most common metrics used to evaluate the predictive performance of a classifier (Smola and Vishwanathan, 2010). The accuracy has a simple idea to show how many of the predictions are done correctly, so, it is given by the number of correct predictions divided by the number of all predictions done. According to Fatourechi et al. (2008), the choice of the right evaluation metric is important as well critical, and depends on characteristics of the application or even on the dataset feature source.

As introduced in section 1, there is a large concern about selecting the best evaluation metric to use for imbalanced domains, and this concern increase when the task is to deal with the multi-class imbalanced domain. In fact, in present times, several metrics have been proposed to apply on this kind of problem (Branco et al., 2017). Choosing the best to use from three different groups (Threshold, Ranking and Probabilistic) has been part of the challenge, once, as presented above, they have different perspectives (Japkowicz, 2013). The point is, in model selection, the classifiers implemented to deal wit multi-class imbalance domain might have different lowest metrics and no single classifier could dominate others inside the group of used metrics, according to the idea presented on Mortaz (2020). Once most of the Threshold and Ranking metrics are based on the confusion matrix, an easier way to evaluate the classification model is to reduce this matrix into one single numeric metric (Prati et al., 2008). Thus, the common metrics like accuracy, recall, precision or F1-score are got based on this idea, and for common metrics used in multi-class imbalanced cases are not different. The largely used on these cases are Accuracy, F1-score, macro and weighted average for precision and recall (Mortaz, 2020).

2.1. Probabilistic metrics for multi-class imbalanced problems

The main purpose of probabilistic metrics, is not related to correct or incorrect class predictions, but on understanding how uncertain is the model in the prediction tasks, giving hard attention to wrong predictions that are highly confident (Jason Brownlee, 2020). In fact, this kind of metrics are in general more concerned to measure the reliability of the classifier, regardless the wrong or right predictions. It is the matter for them, whatever their actions, to verify whether the classifier is getting the information with a good level of certainty or not.

In the process of training some models, there are some probabilistic frameworks used in this process with the main purpose of calibrate their probabilities. The maximum likelihood estimation is a method largely used for this purpose. According to (Ling and Sheng, 2010)

to apply this method to calibrate the classifier, allow the construction of confidence intervals and formal hypothesis testing (Vinayak et al., 2019). It is done in logistic regression, for example, but it is less common in nonlinear classifiers. For this reason, these classifiers are the right candidates to be evaluated with Probabilistic metrics (e.g. SVM or KNN).

The log loss is the common probabilistic metric, for binary cases, given by $LogLoss = -((1-y) * log(1-p(y)) + y * log(p(yi)))$, and it is also known in the literature as cross-entropy in their generalized version, where we can apply for multi-class domain. In fact, cross-entropy is used to determinate or measure the similarity between two distribution functions. Considering $p(x)$ and $q(x)$, the cross entropy is given by Equation 1 for X discrete (only for positive cases), and for continuous Equation, where D_x represent the domains of both functions 2 (Grandini et al., 2020).

$$H(p,q) = -\sum_{D_x} p(x) \log q(x) \tag{1}$$

$$H(p,q) = -\int_{D_x} p(x) \log q(x) \tag{2}$$

In multi-class scenarios, it is considered the Y and $\dot{Y}$, the actual and the predicted discrete variables, assuming each k values, with $K \in \{1, 2, 3...k\}$. Thus, $y^{(i)}$ and $\dot{y}^{(i)}$ are obtained from conditioned random variables of $Y|X$ and $\dot{Y}|X$, and finally, for multi-class classification, the cross-validation is given by Equation 3

$$H(p(y_i), p(\dot{y}_i)) = -\sum_{k=1}^{K} p(Y_i = k|X_i) \log p(\dot{Y}_i = k|X_i) \tag{3}$$

In another hand, there is also the Matthews Correlation Coefficient (MCC), developed by Brian W. Matthews, which is based on Karl Pearson's Coefficient, and the main purpose is comparing the level of agreement between two variables (true labels and predicted ones). The MCC, like most of all probabilistic metrics, works in $[-1, 1]$ interval of continuous and real values, meaning that 1 represent a strong positive correlation between predicted and true values. For multi-class classification, instead of the binary case, the true and the predicted values are given by a sum of each class.

$$MCC = \frac{c \times s - \sum_K p_k \times t_k}{\sqrt{(s^2 - \sum_k p_k^2) * (s^2 - \sum_k t_k^2)}} \tag{4}$$

So, given the multi-class confusion matrix, the MCC metric is given by Equation 4, where $c = \sum_K C_{kk}$ is the total of correctly predicted values, $s = \sum_{ij}^{K} C_{ij}$ is the total of elements, $p_k = \sum_i^K C_{ki}$ represent the number of times that class k was predicted (sum made in column) and $t_k = \sum_K C_{ik}$ represent the number of times that class truly occurred (sum made in row). The MCC is a good indicator for many situations, mainly to indicate the unbalanced prediction model, but its weakness appear in some cases that there are unbalanced results in the model's prediction, where the MCC can show wide fluctuations (Grandini et al., 2020).

The other interesting metric on this group is the Cohen's Kappa, with many MCC similarities for multi-class cases. The main idea on this approach is to measure the concordance between the predicted and the true labels. In fact, Kappa has been largely used in classification as a measure of agreement between observed and predicted or inferred classes of cases in a testing dataset, despite the Delgado and Tibau (2019)'s approach has recommend avoiding use this metric to measure the performance of the classifier. About c, s, p_k, t_k parameter, they have the same functions and values described on the MCC topic.

$$CK = \frac{c \times s - \sum_K p_k \times t_k}{s^2 - \sum_K p_k \times t_k} \tag{5}$$

3. Material and Methods

3.1. Spearman correlation concepts

The Spearman correlation is defined by a non-parametric measure, with the main purpose of evaluating how intense is the relation between two groups of variables and how this intensity can be described as a monotony function. It has, almost, the same idea of Pearson correlation, but Spearman is for variables represented by ranks. This kind of correlation works well with continuous, discrete or even ordinal values. Given X_n and Y_n, a sample of size n, these data have been converted by ranks $R(X_i)$ and $R(Y_i)$ and get Spearman rs by Equation 6.

$$rs = \rho R(X), Y(R) = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \tag{6}$$

Where ρ is the Pearson correlation coefficient, $cov(R(X), R(Y))$ is the covariance but applied for rank data, $\sigma_{R(X)}$ and $\sigma_{R(Y)}$ are the standard deviations of the original variables.

On this work we expect to get ranks composed by distinct integers and for this way, that rs can be computed using the Equation 7, where d_i is the difference between the two ranks we need to compare, given by $d_i = R(X_i) - R(X_i)$

$$rs = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{7}$$

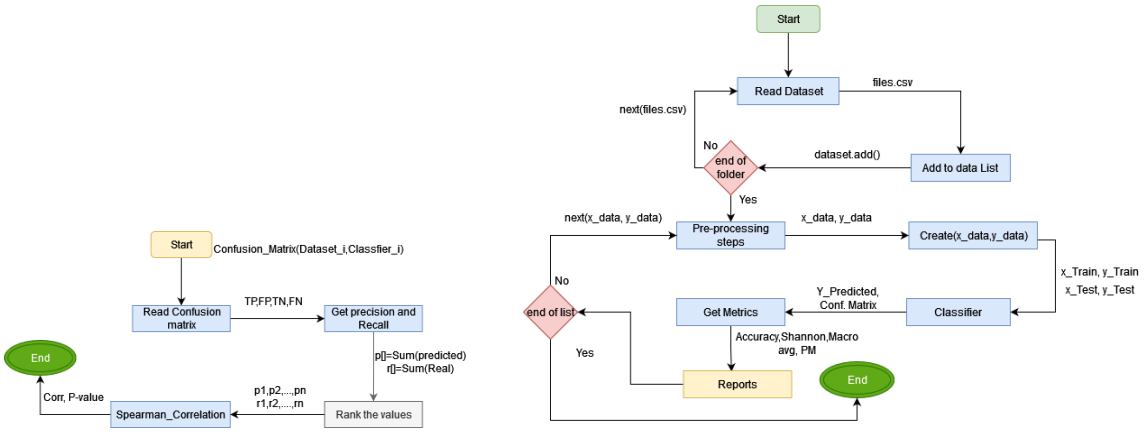
In another hand, after we get the Spearman correlation, it is compulsory to test the evidence, based on the rs value, and check which given hypotheses keep as conclusion. To do this, it is necessary to do the significance test, or calculate the $p - value$ to decide if matter the rs information. For this work, two Hypotheses were considered: the H_0, or null hypotheses, which means the imbalance problem does not affect the way classifier works, and the alternative one H_A, which means that imbalance problem affects the classifier performance.

$$pvalue = r_s \sqrt{\frac{n-2}{1-r_s^2}} \tag{8}$$

3.2. Proposed method

The basic idea of the purposed metric of this work is to use a non-parametric concept, properly the Spearman correlation, to evaluate the imbalanced domains in multi-class cases. The main concern of this method, is not to score the prediction results of a classifier, or even determinate its efficiency. The focus on this idea is, given the results of a classifier, this metric helps to conclude if the predictions (good or not) were affected by the imbalance problem.

The solution was divided by two parts: The main routine, and inside it the specific routine to get our probabilistic metric. The first one is illustrated on 1(b) and the second on 1(a). In the main solution, first we access and read the data folder, and get one by one to charge inside the data structure. After we apply the pre-processing, where we clean the dataset (eliminate index column if it exists and convert all non-numeric attribute into numeric). After, we introduce the selected classifiers to apply each dataset in all of them. Once we generate the confusion matrix, as well as the accuracy, F1-score and Macro average, it is time to get the Proposed Metric (PM), and for this it is used the second part of the solution.



(a) Metric solution (b) General solution

Figure 1: Algorithm of solution

3.3. Datasets

Some datasets were verified, and it was necessary to select them considering different number of attributes, instances and classes. Most of them are from UCI repository, and some from previous works about this topic, developed in papers referred to this work. The table 1 shows, in resume, the features of each dataset used during the experiments

In facts, just looking at the parameters, properly the Shannon Entropy test, it is clear to see that the level of imbalance in a given dataset does not depend on the number of attributes, classes or instances. Thus, there is no dependency between them or at least is a weak dependence, in spite of the level of imbalance commonly is big as big is the number of instances.

Table 1: Resume of datasets features

Dataset	nº Attributes	nº Instances	nº classes	Shannon Entropy
cmc	10	1473	3	0.64
yeast.data	10	1484	10	0.62
reportlost	56	5832	2	0.41
processed.switzerland	14	123	5	0.62
processed.cleveland	14	303	5	0.64
column3C	7	310	3	0.65
floatglass	10	214	6	0.64
zoo	18	101	7	0.66
hayes-roth	6	132	3	0.64

4. Experiments and results

4.1. Preliminary featuring

For these experiments, each dataset was classified by number of attributes, instances and classes. We use the Shannon Entropy (Galar and Kumar, 2017) to measure how imbalanced are the data, and further confront this information with results in classifiers. To interpret the results, based on the PM results and Cross with significance test, they are generated indicators to inform whether the classifier result is affected by the imbalance problem. The initials on the test can be interpreted using the Table 2.

Table 2: Indicators to interpret the cross-validation results

Symbol	Meaning
e	Evidence
ne	Not Evidence
me	midlle Evidence
mne	Middle not Evidence

To run the experiments, we selected two specific groups of classifiers according to their applicability. On first group we selected of simple classifiers with linear and nonlinear features, such as KNN, decision tree, Random forest and logistic regression. In another hand, we select classifiers based on Boosting methods because of their great performance dealing with multi-class problems and good results for imbalanced domains. Inside boosting methods, we selected Gradient Boosting and adaboost using the classifiers of first groups of weak learners.

4.2. Main results

Regarding the results, all algorithms failed the significance test for the "Report Lost" dataset. If we look at the dataset summary, we will see that this particular dataset represents a binary case, and since PM was only built to handle the multi-class case, it cannot

Table 3: Single classifiers experimental results

Dataset	**MCC**	**Accuracy**	**TS**	**Significance**	**Conlusion**
KNN					
cmc	0.35	0.58	0.5	0.67	mne
yeast	0.47	0.59	0.86	0.0	mne
reportlost	0.19	0.65	1.0	n.a	n.a
processedswitz	0.36	0.52	0.67	0.22	me
processedClev	0.16	0.56	0.9	0.04	mne
column3C	0.75	0.84	-1.0	0.0	e
floatglass	0.6	0.72	-0.12	0.82	mne
zoo	0.93	0.95	0.0	1.0	ne
hayes	0.36	0.52	0.0	1.0	ne
Decicion tree					
cmc	0.26	0.52	0.5	0.67	mne
yeast	0.37	0.5	-0.02	0.95	mne
reportlost	0.16	0.61	1.0	n.a	n.a
processedswitz	0.18	0.4	0.67	0.22	mne
processedClev	0.27	0.52	-0.1	0.87	mne
column3C	0.65	0.77	-0.5	0.67	mne
floatglass	0.62	0.72	-0.71	0.12	mne
zoo	0.93	0.95	0.0	1.0	ne
hayes	0.69	0.74	-0.87	0.33	me
Random Forest					
cmc	0.29	0.54	0.5	0.67	mne
yeast	0.48	0.59	0.68	0.03	mne
reportlost	0.21	0.65	1.0	n.a	n.a
processedswitz	0.06	0.32	0.67	0.22	me
processedClev	0.34	0.62	0.7	0.19	me
column3C	0.74	0.84	-1.0	0.0	e
floatglass	0.67	0.77	-0.54	0.27	mne
zoo	0.93	0.95	0.0	1.0	ne
hayes	0.69	0.74	-0.87	0.33	me
Logistic Regression					
cmc	0.24	0.51	0.5	0.67	mne
yeast	0.42	0.55	0.89	0.0	mne
reportlost	0.21	0.65	1.0	n.a	n.a
processedswitz	0.24	0.44	0.87	0.05	mne
processedClev	0.37	0.64	0.7	0.19	me
column3C	0.75	0.84	-1.0	0.0	e
floatglass	0.54	0.67	-0.71	0.12	me
zoo	1.0	1.0	n.a	n.a	me
hayes	0.34	0.41	-0.87	0.33	me

get any definition in conclusion. Incidentally, it was one of the warnings used in our research to test the delimitation of the scope of work.

On Table 3, were explored single classifiers (linear and non-linear), to get answer about the PM metric developed on this work, accuracy as a common Threshold metric, and MCC. In common, for all datasets we see that the "e" result is linked with high MCC coefficient values and all occurrences were with "columm3C" dataset, and this observation is complemented with "ne" result that present also high MCC (and Accuracy) for all situations and most cases happen with "hayes" dataset. Another attention is necessary for the "Zoo" dataset, because almost all classifiers got the "ne" result and the numeric metrics value got little variation during the experiments.

About the effect on the results of classifier, there are many situations classified as middle (or middle not) evidence, which means that this classifier with the dataset charged, can or not be affected by the imbalance problem (like 50/50). For singular ones, decision tree did not get evidence in all cases, so its configure a good candidate to work with more powerful experiments because this classifier probably is less affected by the imbalanced problem. Actually, this fact can be verified in Random forest, where we have the large number of middle not ("mne") and not("ne") evidence. Thus, this behaves can be also verified on Ada boost method, with decision tree as weak classifier. in another hand, the KNN was the one with more "not evidence" and "middle not evidence" results (Table 3). So it's mean, for this preliminary study, the KNN was the one with more "not evidence" and "middle not evidence" results (Table 3). So, for this preliminary study, the KNN might be the classifier less affected by imbalanced problem (not necessarily with positive or negative score).

4.3. Future Work

Nowadays, we have a large bibliography about methods to deal and evaluate multi-class imbalanced problems, and PM is one more contribution for this research area. For the future, we expect to use the PM to evaluate a bigger number of classifiers with high potential to deal with multi-class imbalance problems, and explore it in more datasets, as well as develop a comparison study with other probabilistic methods and explore the relation between them and threshold metrics.

5. Conclusion

On this work, we propose a probabilistic metric (PM) for the multi-class imbalanced domain, which main concern is to measure the effect of the imbalance in the classifiers' results. A brief approach were done in section II to get clear the kind of metrics and challenges about it that can be used in multi-class imbalanced domains, and more specifically to probabilistic metrics, which is the main focus of the work. For the experiments, nine datasets from different areas were selected, and we used to test the effect of the MP in two groups of classifiers: the singular ones and the Boosting methods. In general, we could observe that, for imbalanced cases, the threshold metrics are not good to evaluate the prediction of classifiers, once the results can be strongly affected by the distribution of the classes. In another hand, MCC showed some relation with PM conclusions.

Table 4: Boosting classifiers experimental results

Dataset	**MCC**	**Accuracy**	**PM**	**Significance**	**Conlusion**
Gradient Boosting					
cmc	0.33	0.56	-0.5	0.67	mne
yeast	0.23	0.4	0.63	0.05	mne
reportlost	0.24	0.66	1.0	n.a	n.a
processedswitz	-0.16	0.2	0.87	0.05	mne
processedClev	0.2	0.48	-0.3	0.62	mne
column3C	0.72	0.82	-1.0	0.0	e
floatglass	0.62	0.72	-0.75	0.08	me
zoo	0.93	0.95	0.0	1.0	ne
hayes	0.72	0.78	-0.87	0.33	me
ADA+DT					
cmc	0.31	0.55	0.5	0.67	mne
yeast	0.25	0.41	0.22	0.54	mne
reportlost	0.22	0.66	1.0	n.a	n.a
processedswitz	0.2	0.4	0.3	0.62	mne
processedClev	0.29	0.52	0.1	0.87	mne
column3C	0.44	0.58	0.5	0.67	mne
floatglass	0.11	0.3	0.13	0.8	mne
zoo	0.74	0.81	0.26	0.62	mne
hayes	0.49	0.44	-0.87	0.33	me
ADA+ExtraTree					
cmc	0.29	0.54	0.5	0.67	mne
yeast	0.29	0.44	-0.03	0.93	mne
reportlost	0.15	0.61	1.0	n.a	n.a
processedswitz	0.03	0.28	0.1	0.87	mne
processedClev	0.27	0.52	-0.1	0.87	mne
column3C	0.46	0.66	-0.5	0.67	mne
floatglass	0.6	0.72	0.31	0.55	mne
zoo	1.0	1.0	n.a	n.a	n.a
hayes	0.68	0.78	-0.87	0.33	me
ADA+LR					
cmc	0.24	0.51	0.5	0.67	mne
yeast	0.06	0.31	0.7	0.02	mne
reportlost	0.21	0.65	1.0	n.a	n.a
processedswitz	-0.02	0.2	-0.7	0.19	me
processedClev	0.36	0.61	-0.3	0.62	mne
column3C	0.73	0.82	-0.5	0.67	mne
floatglass	0.17	0.42	-0.46	0.35	mne
zoo	0.93	0.95	0.0	1.0	ne
hayes	0.24	0.41	-0.87	0.33	me

Acknowledgment

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project

References

Paula Branco, Luís Torgo, and Rita P. Ribeiro. Relevance-based evaluation metrics for multi-class imbalanced domains. In Jinho Kim, Kyuseok Shim, Longbing Cao, Jae-Gil Lee, Xuemin Lin, and Yang-Sae Moon, editors, *Advances in Knowledge Discovery and Data Mining*, pages 698–710, Cham, 2017. Springer International Publishing.

Subhasish Deb, Arup Kumar Goswami, Rahul Lamichane Chetri, and Rajesh Roy. Prediction of plug-in electric vehicle's state-of-charge using gradient boosting method and random forest method. In *2020 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES)*, pages 1–6, 2020. doi: 10.1109/PEDES49360.2020.9379906.

Rosario Delgado and Xavier-Andoni Tibau. Why cohen's kappa should be avoided as performance measure in classification. *PLoS ONE*, 14(9):e0222916–e0222916, 2019. ISSN 1932-6203.

Mehrdad Fatourechi, Rabab K. Ward, Steven G. Mason, Jane Huggins, Alois Schlögl, and Gary E. Birch. Comparison of evaluation metrics in classification applications with imbalanced datasets. In *2008 Seventh International Conference on Machine Learning and Applications*, pages 777–782, 2008.

Diego Galar and Uday Kumar. Chapter 3 - preprocessing and features. In Diego Galar and Uday Kumar, editors, *eMaintenance*, pages 129–177. Academic Press, 2017. ISBN 978-0-12-811153-6. URL https://www.sciencedirect.com/science/article/pii/B9780128111536000038.

Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview, 2020. URL https://arxiv.org/abs/2008.05756.

Qiong Gu, Li Zhu, and Zhihua Cai. Evaluation measures of the classification performance of imbalanced data sets. In Zhihua Cai, Zhenhua Li, Zhuo Kang, and Yong Liu, editors, *Computational Intelligence and Intelligent Systems*, pages 461–471, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-04962-0.

Nathalie Japkowicz. *Assessment Metrics for Imbalanced Learning*, pages 187–206. 06 2013. ISBN 9781118074626.

Jason Brownlee. Tour of evaluation metrics for imbalanced classification, 2020.

Piyasak Jeatrakul, Kok Wong, and Chun Fung. Classification of imbalanced data by combining the complementary neural network and smote algorithm. pages 152–159, 11 2010. ISBN 978-3-642-17533-6.

Kamaldeep Kaur, Jasmeet Kaur Name, and Jyotsana Malhotra. Evaluation of imbalanced learning with entropy of source code metrics as defect predictors. In *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, pages 403–409, 2017.

Charles X. Ling and Victor S. Sheng. *Cost-Sensitive Learning*, pages 231–235. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. URL https://doi.org/10.1007/978-0-387-30164-8_181.

Ebrahim Mortaz. Imbalance accuracy metric for model selection in multi-class imbalance classification problems. *Knowledge-Based Systems*, 210:106490, 2020. ISSN 0950-7051. URL https://www.sciencedirect.com/science/article/pii/S0950705120306195.

Ronaldo Cristiano Prati, Gustavo Enrique Almeida Prado Alves Batista, and Maria Carolina Monard. Evaluating classifiers using roc curves. *IEEE Latin America Transactions*, 6(2):215–222, 2008.

Alex Smola and S.V.N. Vishwanathan. *Introduction to Machine Learning*. press syndicate of the university of cambridge, 2010.

Yiming Tian and Xitai Wang. Svm ensemble method based on improved iteration process of adaboost algorithm. In *2017 29th Chinese Control And Decision Conference (CCDC)*, pages 4026–4032, 2017. doi: 10.1109/CCDC.2017.7979205.

Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham Kakade. Maximum likelihood estimation for learning populations of parameters. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6448–6457. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/vinayak19a.html.

Cong-man Wang, Hui-zhi Yang, Fa-chao Li, and Rui-xue Fu. Two stages based adaptive sampling boosting method. In *2006 International Conference on Machine Learning and Cybernetics*, pages 2925–2927, 2006.

Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022.

Ni Wayan Surya Wardhani, Masithoh Yessi Rochayani, Atiek Iriany, Agus Dwi Sulistyono, and Prayudi Lestantyo. Cross-validation metrics for evaluating classification performance on imbalanced data. In *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pages 14–18, 2019. doi: 10.1109/IC3INA48034.2019.8949568.

Ming Zheng, Tong Li, Rui Zhu, Yahui Tang, Mingjing Tang, Leilei Lin, and Zifei Ma. Conditional wasserstein generative adversarial network-gradient penalty-based approach to alleviating imbalanced data classification. *Information Sciences*, 512:1009–1023, 2020. ISSN 0020-0255.