Evaluating the Prediction Bias Induced by Label Imbalance in Multi-label Classification

Luca Piras
Data Science and Big Data Analytics, EURECAT (Centre Tecnológic de Catalunya)
Barcelona, Spain
luca.piras@eurecat.org

Ludovico Boratto
Department of Mathematics and Computer Science, University of Cagliari
Cagliari, Italy
ludovico.boratto@acm.org

Guilherme Ramos
Dept. of Electrical and Computer Engineering - University of Porto
Porto, Portugal
guilhermeramos21@gmail.com

ABSTRACT

Prediction bias is a well-known problem in classification algorithms, which tend to be skewed towards more represented classes. This phenomenon is even more remarkable in multi-label scenarios, where the number of underrepresented classes is usually larger. In light of this, we hereby present the *Prediction Bias Coefficient (PBC)*, a novel measure that aims to assess the bias induced by label imbalance in multi-label classification. The approach leverages Spearman's rank correlation coefficient between the label frequencies and the F-scores obtained for each label individually. After describing the theoretical properties of the proposed indicator, we illustrate its behaviour on a classification task performed with state-of-the-art methods on two real-world datasets, and we compare it experimentally with other metrics described in the literature.

CCS CONCEPTS

• **Information systems → Clustering and classification**; **Evaluation of retrieval results**.

KEYWORDS

classification bias, multi-label classification, imbalance, evaluation

ACM Reference Format:
Luca Piras, Ludovico Boratto, and Guilherme Ramos. 2021. Evaluating the Prediction Bias Induced by Label Imbalance in Multi-label Classification. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482100

1 INTRODUCTION

Classification algorithms are known to suffer from the *prediction bias* issue, as the imbalance between the classes causes them to penalise classes that appear less frequently in the training set [1, 2]. This behaviour is more pronounced in multi-label settings, in which each instance can be associated with a variable number of labels [3]. The interdependence of co-occurring labels and the larger number of underrepresented labels makes the learning process in this scenario remarkably more complex [4, 5]. This challenge is particularly relevant in the current technological era, in which increasingly larger amounts of multi-medial Big Data are being annotated with tags or categories for a variety of tasks [6, 7]. Moreover, there are applications such as job recommendations, credit scoring, or fraud detection, where prediction bias can potentially translate into the discrimination or exclusion of certain minorities of people [8]. In this sense, we can include the prediction bias problem in the field of *algorithmic fairness*, which is receiving more and more attention in Information Retrieval (IR) and Machine Learning (ML) [9].

In the last decades, a wide range of approaches have been proposed to cope with the label imbalance in multi-label classification [10–12]. A common strategy consists in designing *algorithmic adaptations* of well-known classification techniques [13–15]. *Sampling* methods are an alternative, which operates directly on the data to obtain a balanced dataset on which to train the classifier [4, 16]. Typically, this is achieved either by discarding instances of the majority classes (undersampling [17]) or adding new instances of the minority classes (oversampling [18]). Moreover, researchers have proposed measures that allow evaluating and/or optimise a multi-label classifier potentially affected by imbalance [16, 19, 20]. This is the research branch that inspired the indicator described in this paper and on which, therefore, we will focus our attention.

Several metrics have been proposed to measure the imbalance degree of a dataset [4, 19, 21, 22]. The *Imbalance Ratio per Label (IRLbl)* measures the ratio between the frequencies of the majority label and a given one. The *Mean Imbalance Ratio (MeanIR)* is an indicator of the average level of imbalance between the classes, obtained as the mean of the IRLbl scores, while the *Coefficient of Variation of the IRLbl (CVIR)* indicates how much the level of imbalance differs among the labels. What all these indicators have in common is that they aim to evaluate the composition of a dataset, *independently* of the performance of a classifier applied on those data. In other words, these metrics gauge the imbalance by taking into account solely the distribution of the labels, while ignoring how imbalance affects the correctness of the classification task.

Standard metrics from IR and ML are also commonly employed to evaluate the output of a multi-label classifier [11]. It is well-known that precision, recall, and F-score are to be preferred w.r.t. common accuracy in imbalanced scenarios because the latter is not sensitive to data distribution[23]. F-score, obtained as the weighted harmonic mean of precision and recall, is particularly suited as a global indicator of the functionality of a classifier [20, 24]. Even

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482100

though the F-score preserves the sensitivity to the data distribution, it fails to capture whether the classifier discriminates against certain labels and to which extent the classifier's performance is affected by class imbalance. The same holds for all other evaluation techniques, including ROC curves, precision-recall curves, and cost curves [11].

In light of our analysis of the literature, we propose the *Prediction Bias Coefficient (PBC)*, an indicator that measures the strength of the correlation between the label imbalance in a dataset and the performance obtained by a multi-label classifier trained on the same dataset. The approach exploits the *Spearman's rank correlation coefficient* between the label frequencies and the F-scores obtained for each label individually. In terms of algorithmic fairness, we can interpret the PBC indicator as a way of capturing the intensity of the classifier's discrimination against the minority classes.

The contributions of our work can be summarised as follows: (i) we propose a novel evaluation indicator that measures the correlation between the label imbalance and the classification performance (Section 2); (ii) we visually illustrate the proposed metric on a multi-label classification task carried out with state-of-the-art techniques on two real-world datasets (Section 3); (iii) we compare experimentally the proposed indicator with other existing metrics (Section 3).

2 THE PREDICTION BIAS COEFFICIENT

In this section, we formally define the proposed indicator. Then, we proceed to discuss its interpretation and its utility.

2.1 Definition

We denote sets by uppercase letters (e.g., A), elements of sets by lowercase letters (e.g., a), and vectors by bold lowercase letters (e.g., $\mathbf{a}$). Moreover, we use $|A|$ to denote the number of elements in A. A *dataset* is a pair of sets (X, Y), where X is a set of observations and Y is a set of ground truth labels, such that $|X| = |Y|$. If $\mathcal{X}$ is the n-dimensional space such that $X \subseteq \mathcal{X}$, we call $\mathcal{X}$ the *feature space*.

A multi-label classification task with a dataset (X, Y) and a set of labels L is given, with $|L| \gg 2$; each instance in (X, Y) is a pair $(\mathbf{x}, y)$, where $\mathbf{x}$ is a vector in a feature space $\mathcal{X}$ associated with the ground truth variable-length list of labels y, such that $y \subseteq L$, where $0 \leq |y| \leq |L|$. The data are split between a training set (X_{train}, Y_{train}) and a test set (X_{test}, Y_{test}), i.e., $(X_{train}, Y_{train}) \cup (X_{test}, Y_{test}) = (X, Y)$. The label l frequency, $freq_l$, is estimated as the proportion of label lists in Y_{train} that contain l; formally:

$$freq_l = \frac{|\ \{y \mid l \in y, \forall y \in Y_{train}\}\ |}{|\ Y_{train}\ |}, \forall l \in L. \tag{1}$$

We formalise a classifier as a function mapping elements of the feature space into elements of the label space: $h(\cdot\,; \theta) : \mathcal{X} \mapsto \wp(L)$, where for a set X, $\wp(X)$ denotes the non-empty parts of X, and θ are the parameters of the classifier. After being trained on (X_{train}, Y_{train}), the classifier is tested on (X_{test}, Y_{test}), thus producing W, a set of lists w containing the predicted labels for each instance in the test set ($w \subseteq L$, with $0 \leq |w| \leq |L|$). The binary F-score can be computed separately for each label l by checking the ground truth lists and the prediction lists that contain l; formally:

$$F_l = \frac{TP_l}{TP_l + \frac{1}{2}(FP_l + FN_l)}, \forall l \in L, \tag{2}$$

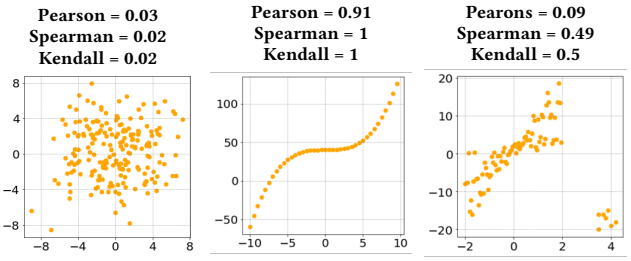


Fig. 1: Comparison between Pearson's, Spearman's and Kendall's correlations for: a) roughly elliptically distributed variables; b) perfectly monotonic non-linear relationship; c) distribution with prominent outliers.

where $TP_l =| \{(y, w) \mid l \in y \wedge l \in w, \forall y \in Y_{test}, \forall w \in W\} |$ are the true positives for label l, $FP_l =| \{(y, w) \mid l \notin y \wedge l \in w, \forall y \in Y_{test}, \forall w \in W\} |$ are the false positives and $FN =| \{(y, w) \mid l \in y \wedge l \notin w, \forall y \in Y_{test}, \forall w \in W\} |$ are the false negatives.

Finally, let rg_{freq} and rg_F denote the ranks of $freq$ and F, respectively (the rank of a list of values maps each value to its positional index in the decreasingly sorted list, with indexes starting from 1). We now have all the elements to define formally the PBC, calculated as the Spearman's rank correlation coefficient between the proportion variables $freq$ and F.

$$PBC = \frac{cov(rg_{freq}, rg_F)}{\sigma_{rg_{freq}} \sigma_{rg_F}}, \tag{3}$$

where cov is the covariance between two random variables, and σ is the standard deviation.

The Spearman's correlation [25] is a non-parametric indicator that measures the statistical dependence between the rankings of two variables. It does so by assessing how well the relationship between the two variables can be described in terms of a monotonic function. Its ranges between -1 and 1, which indicate a perfect opposite and direct correlation, respectively, while 0 indicates no dependence [26]. The Spearman's correlation between two variables is equal to the Pearson's correlation between the rank values of those two variables [25, 27]. When two variables are roughly elliptically distributed without strong outliers, Spearman and Pearson give similar values (Fig. 1a). However, the Pearson's coefficient is limited to the assessment of *linear* relationship and reaches value 1 only when the two random variables are perfectly linearly related, which would be too strict a condition for our application (Fig. 1b). On top of that, Spearman's correlation is more robust to prominent outliers compared to Pearson's coefficient, because it limits the outliers to the values of their rank (Fig. 1c). For these reasons, the choice fell on the Spearman's coefficient when designing the new Prediction Bias Coefficient. We refer the reader to [28] for an analysis of the sample size requirements for estimating the Spearman's and Pearson's correlations. Spearman's measure could possibly be replaced by other rank correlation metrics, such as Kendall's τ coefficient; Fig. 1 shows a comparison with the other coefficients.

2.2 Interpretation

The PBC keeps the properties of the Spearman's coefficient, as it is positive if the F-scores on single labels tend to increase together with the label frequencies. In particular, it reaches its maximum value (1) whenever these two quantities are perfectly monotonically related. This scenario would imply that the classifier is strongly biased towards majority labels, penalising the underrepresented ones. Conversely, if higher label frequencies are associated with lower F-scores, the PBC is negative. It reaches the lowest value (-1) whenever those two quantities are related by a perfectly monotonic decreasing function. This minimum would happen in the scenario (atypical but theoretically possible) in which the algorithm consistently favours the minority labels at the expense of the most frequent ones. The ideal classifier, in terms of bias and fairness, would be the one that achieves PBC scores tending to 0, meaning that label frequencies and label-wise F-scores are uncorrelated.

To show the advantage of PBC, let us consider an example. Suppose that a dataset has five labels $l1$, $l2$, $l3$, $l4$, $l5$, with frequencies 0.1, 0.2, 0.3, 0.4, 0.5, respectively. Let us also suppose that a classifier $h1$ achieves on these labels the following respective F-scores: 0.5, 0.3, 0.4, 0.2, 0.6. The macro-averaged F-score would be 0.4, while the PBC would equal 0.1. On the other hand, let us suppose that a different classifier $h2$ achieves, on the same labels, the following F-scores: 0.2, 0.3, 0.4, 0.5, 0.6 (please note that they are the same values obtained by $h1$, but in an order that strictly follows the label frequencies). In this case, the macro-averaged F-score would still be 0.4, like for $h1$, but the PBC would equal 1, indicating that $h2$ is heavily biased towards most represented classes. This demonstrates that the F-score fails to detect how different classifiers are influenced in different ways by the dataset imbalance, while this information is effectively conveyed by the PBC.

We notice that, in the definition of the PBC, the F-score can be substituted with the label-wise precision or recall, depending on the desired perspective. We decided to present the formulation based on F-score because this metric encompasses the other two. It is important to clarify that the PBC does not convey any information on the classification correctness, but only on the degree of its dependency with the imbalance. For instance, a classifier that achieves similarly low F-scores for all labels would get a better PBC compared to a highly accurate model that penalises minority classes. For this reason, the PBC should be considered as a complementary indicator w.r.t. standard accuracy metrics.

Interestingly, a random classifier that outputs each label with the same probability p would get a PBC that tends asymptotically to 1. In fact, it can be proven that, in such scenario: (i) the precision obtained on label l tends asymptotically to the frequency of l in the dataset; (ii) the recall tends for all labels to the same value, which is the probability p; (iii) the F-score, which is the harmonic mean of precision and recall, is monotonically correlated with the precision scores and, consequently, with the label frequency.

Finally, we observe that we can also apply the PBC to multi-class scenarios, where (unlike the multi-label case) each instance is associated with exactly one label from a set of n labels, with $n > 2$ [29]. However, we describe and evaluate the metric in the multi-label scenario because the latter generalises the former one.

Table 1: Statistics about the two datasets employed in the experiments. |X| is the total number of samples, |L| the total number of labels, "card" is the cardinality, "dens" is the density and "% labeled" is the percentage of samples with at least one label.

dataset	|X|	|L|	card	dens	% labeled
Reuters-21578	19,043	119	0.69	0.01	54%
Webscope-R4	106,959	16	0.58	0.03	55%

3 EXPERIMENTAL EVALUATION

We evaluate the proposed measure in a multi-label classification task carried out with state-of-the-art algorithms on two real-world datasets. In addition, we visually highlight the correlation detected by the PBC and compare the PBC against other evaluation metrics. Our focus is on the metrics and their application in an imbalanced multi-label/multi-class classification scenario, so the classifier can be seen as a black box that can be arbitrarily changed. Hence, our approach is generalizable to any multi-label classification algorithm.

3.1 Methodology

We developed our experimental framework using the Python language with standard modules, such as numpy[1], scipy[2] and scikit-learn[3]. The datasets employed in our analysis are *Reuters-21578*[4], consisting of news stories tagged with a list of economic categories, and *Webscope-R4*[5], which includes a corpus of movie synopses associated to one or more genres. For our experiments, we selected all samples associated to at least 1 label. Table 1 shows statistics about the two resulting datasets. The *cardinality* is defined as the average number of labels per instance, while the *density* is given by the cardinality divided by the total number of labels [16].

We split each dataset into a series of training and test sets, using a 10-fold cross-validation approach. A multi-label text classifier was trained exploiting the `TextCategorizer` model provided by the Python library spacy (version 2.x)[6]. This algorithm uses a stacked ensemble of bag-of-words and a Convolutional Neural Network [30], aligned with the state-of-the-art [31, 32]. Since the classifier's output is a list of probabilities over all possible labels, we selected the predicted labels using a threshold $t = 0.5$, interpreted as a confidence level. We set the value for t experimentally.

For each test set, together with the PBC, we measured the aforementioned MeanIR and CVIR (which describe the label imbalance without taking into account the classification correctness [21]), the balanced accuracy [33] and the macro-averaged F-score (which, in multi-label scenarios, is more suited than the micro-averaged version [20]). Additionally, we render a scatter-plot to visually inspect the correlation between the label imbalance and the performance.

3.2 Results

Table 2 summarises the results obtained on the two datasets. The most relevant finding is that the dataset associated with a higher degree of imbalance (namely, *Webscope-R4*, as pointed out by MeanIR and CVIR) is also the one on which the classifier achieves, on one hand, a *lower* value for both the balanced accuracy and the macro-averaged F-score, and, on the other hand, a *higher* value of PBC. This suggests that the latter metric can capture the relation between

[1] https://numpy.org/ [2] https://www.scipy.org/ [3] https://scikit-learn.org/
[4] http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection
[5] https://webscope.sandbox.yahoo.com/catalog.php?datatype=r [6] https://spacy.io/

Table 2: Results obtained on the two datasets. The shown values are the averages ± standard deviations of the values obtained on the 10 folds of the cross-validation.

dataset	MeanIR	CVIR	B. Acc.	F-score	PBC
Reuters	829.11 ± 37.3	1.34 ± 0.02	0.76 ± 0.03	0.51 ± 0.07	0.53 ± 0.1
Webscope	2757.94 ± 609	2.16 ± 0.47	0.69 ± 0.03	0.44 ± 0.06	0.66 ± 0.19

the imbalance degree and the performance. The visual inspection offered by the two scatter-plots of Fig. 2, which illustrate the correlation between label frequencies and label-wise F-scores, allows us observe that, in *Webscope-R4*, the macro-averaged F-score is dragged down by the very poor accuracy obtained on severely underrepresented labels (visible in the bottom-left corner of Fig. 2b). On the other hand, in *Reuters-21578* (Fig. 2a), the low F-scores obtained on very infrequent labels (bottom-left corner) are balanced by very high scores obtained on equally infrequent labels (upper-left corner), thus leading to a lower PBC compared to *Webscope-R4*.

The label distribution in Fig. 2a might induce readers to believe that the PBC is well above 0 because of the strong influence of the two outlier labels (namely, 'acq' and 'earn'). However, this is not the case. Indeed, to further demonstrate that the PBC inherits from the Spearman's coefficient its robustness to prominent outliers, we re-computed the PBC on *Reuters-21578* without taking into account the scores obtained on those two labels and we obtained still the same PBC value (0.53). Fig. 3 can be thought of as a zoom-in of the left-hand side of Fig. 2a; it shows that even in the scenario without outliers, there is quite a pronounced dependence between the label frequencies and the F-scores, which was not evident in Fig. 2a because of the horizontal deformation of the image. (The red trend line in Figs. 2 and 3, obtained as a 1st-degree polynomial interpolation of the points, is shown for illustration purpose only; although it gives a useful geometric intuition of the correlation, it is not strictly related to the PBC from a formal standpoint).

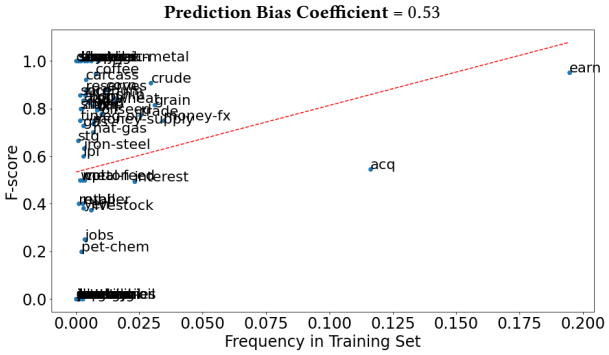
We can note that the PBC values obtained on different portions of the same dataset can vary significantly (as shown by the high standard deviations in Table 2). For this reason, we recommend a cross-validation of this metric, to make the evaluation more robust.

4 CONCLUSIONS

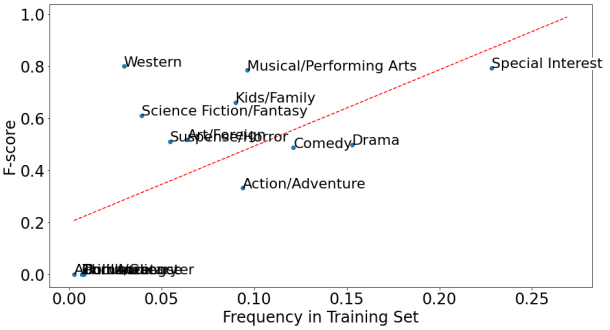
In this paper, we presented the *Prediction Bias Coefficient* (PBC), an indicator that measures the correlation between the label imbalance and the correctness of a multi-label classifier. This metric allows assessing the bias towards majority classes, thus providing a tool to evaluate a classifier beyond accuracy. Potentially, it can be employed in the context of *algorithmic fairness* to estimate to which extent an algorithm discriminates against underrepresented categories.

After giving a formal definition, we provided empirical evidence of its utility on two datasets, in a text classification task performed with state-of-the-art methods. The PBC effectively captures the correlations while being complementary to metrics that gauge only the imbalance of a dataset (Mean Imbalance Ratio and Coefficient of Variation of the Imbalance Ratio) or only the correctness (F-score).

In the future, we aim to include more datasets into the experimental analysis; by using different types of data (images, audio, time-series, etc.) and a wider variety of label distributions, we would analyze in more depth the behaviour of the proposed metric. To conclude, we intend to test the PBC on classification algorithms that are specifically designed to cope with label imbalance (via either resampling methods or algorithmic adaptations), with the goal of observing how sensitive our indicator is to such strategies and how it can help in finding the most effective one.



(a) Reuters-21578



(b) Webscope-R4

Fig. 2: Correlation between the frequency of the labels in the training set and the label-wise F-scores. The red line, obtained as a 1st-degree polynomial interpolation of the points, indicates the trend.

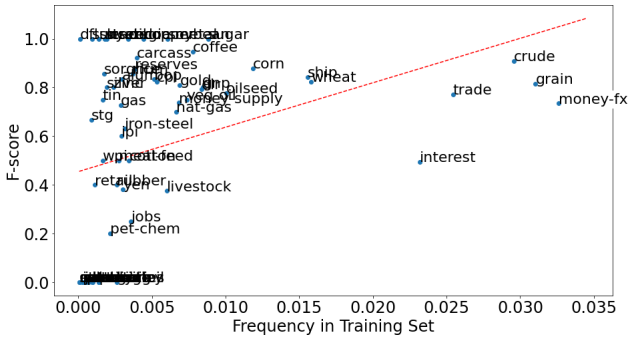


Fig. 3: Correlation between label frequencies and F-scores in the *Reuters-21578* dataset, without the two outlier labels with highest frequency ('acq' and 'earn').

Reproducibility. The code implementing our metric and experimental framework are available at https://github.com/luca-24/prediction-bias-coefficient.

Acknowledgments. This work was supported in part by project AVALUA, funded by ACCIÓ, and by project RELIABLE (PTDC/EEI-AUT/3522/2020) funded by FCT/MCTES.

REFERENCES

[1] Haibo He and Yunqian Ma. Imbalanced learning: foundations, algorithms, and applications. 2013.

[2] Joffrey L Leevy, Taghi M Khoshgoftaar, Richard A Bauder, and Naeem Seliya. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):1–30, 2018.

[3] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[4] Bin Liu, Konstantinos Blekas, and Grigorios Tsoumak. Multi-label sampling based on local label imbalance. *arXiv preprint arXiv:2005.03240*, 2020.

[5] Min-Ling Zhang, Yu-Kun Li, Hao Yang, and Xu-Ying Liu. Towards class-imbalance aware multi-label learning. *IEEE Transactions on Cybernetics*, 2020.

[6] Yu Zhang, Yin Wang, Xu-Ying Liu, Siya Mi, and Min-Ling Zhang. Large-scale multi-label classification using unknown streaming images. *Pattern Recognition*, 99:107100, 2020.

[7] Fangfang Luo, Wenzhong Guo, Yuanlong Yu, and Guolong Chen. A multi-label classification algorithm based on kernel extreme learning machine. *Neurocomputing*, 260:313–320, 2017.

[8] Deborah Hellman. Measuring algorithmic fairness. *Va. L. Rev.*, 106:811, 2020.

[9] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27, 2018.

[10] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

[11] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[12] Li Li and Houfeng Wang. Towards label imbalance in multi-label classification with many labels. *arXiv preprint arXiv:1604.01304*, 2016.

[13] Alberto Fernández, Victoria López, Mikel Galar, MaríA José Del Jesus, and Francisco Herrera. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-based systems*, 42:97–110, 2013.

[14] Zachary Daniels and Dimitris Metaxas. Addressing imbalance in multi-label classification using structured hellinger forests. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[15] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 647–657, 2019.

[16] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:3–16, 2015.

[17] Muhammad Atif Tahir, Josef Kittler, and Fei Yan. Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recognition*, 45(10):3738–3750, 2012.

[18] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Mlsmote: approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385–397, 2015.

[19] Jonathan Ortigosa-Hernández, Inaki Inza, and Jose A Lozano. Measuring the class-imbalance extent of multi-class problems. *Pattern Recognition Letters*, 98:32–38, 2017.

[20] Ignazio Pillai, Giorgio Fumera, and Fabio Roli. Designing multi-label classifiers that maximize f measures: State of the art. *Pattern Recognition*, 61:394–404, 2017.

[21] Francisco Charte, Antonio Rivera, María José del Jesus, and Francisco Herrera. A first approach to deal with imbalance in multi-label datasets. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 150–160. Springer, 2013.

[22] Rui Zhu, Ziyu Wang, Zhanyu Ma, Guijin Wang, and Jing-Hao Xue. Lrid: A new metric of multi-class imbalance degree based on likelihood-ratio test. *Pattern Recognition Letters*, 116:36–42, 2018.

[23] Marcus A Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML-2003 workshop on learning from imbalanced data sets II*, volume 2, pages 2–1, 2003.

[24] Krzysztof Dembczynski, Arkadiusz Jachnik, Wojciech Kotlowski, Willem Waegeman, and Eyke Hüllermeier. Optimizing the f-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In *International conference on machine learning*, pages 1130–1138. PMLR, 2013.

[25] Charles Spearman. The proof and measurement of association between two things. 1961.

[26] Jerrold H Zar. Spearman rank correlation. *Encyclopedia of biostatistics*, 7, 2005.

[27] Joost CF de Winter, Samuel D Gosling, and Jeff Potter. Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological methods*, 21(3):273, 2016.

[28] Douglas G Bonett and Thomas A Wright. Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*, 65(1):23–28, 2000.

[29] Mohamed Aly. Survey on multiclass classification methods. *Neural Netw*, 19:1–9, 2005.

[30] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[31] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[32] Marcin Michał Mirończuk and Jarosław Protasiewicz. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, 106:36–54, 2018.

[33] Lawrence Mosley. A balanced approach to the multi-class imbalance problem. 2013.