PLOS ONE

RESEARCH ARTICLE

Why Cohen's *Kappa* should be avoided as performance measure in classification

Rosario Delgado[1☯]*, **Xavier-Andoni Tibau**[2☯]

1 Department of Mathematics, Universitat Autònoma de Barcelona, Campus de la UAB, Cerdanyola del Vallès, Spain, **2** Advanced Stochastic Modelling research group, Universitat Autònoma de Barcelona, Campus de la UAB, Cerdanyola del Vallès, Spain

☯ These authors contributed equally to this work.
* delgado@mat.uab.cat

Abstract

We show that Cohen's *Kappa* and Matthews Correlation Coefficient (MCC), both extended and contrasted measures of performance in multi-class classification, are correlated in most situations, albeit can differ in others. Indeed, although in the symmetric case both match, we consider different unbalanced situations in which *Kappa* exhibits an undesired behaviour, i.e. a worse classifier gets higher *Kappa* score, differing qualitatively from that of MCC. The debate about the incoherence in the behaviour of *Kappa* revolves around the convenience, or not, of using a relative metric, which makes the interpretation of its values difficult. We extend these concerns by showing that its pitfalls can go even further. Through experimentation, we present a novel approach to this topic. We carry on a comprehensive study that identifies an scenario in which the contradictory behaviour among MCC and *Kappa* emerges. Specifically, we find out that when there is a decrease to zero of the entropy of the elements out of the diagonal of the confusion matrix associated to a classifier, the discrepancy between *Kappa* and MCC rise, pointing to an anomalous performance of the former. We believe that this finding disables *Kappa* to be used in general as a performance measure to compare classifiers.

Citation: Delgado R, Tibau X-A (2019) Why Cohen's *Kappa* should be avoided as performance measure in classification. PLoS ONE 14(9): e0222916. https://doi.org/10.1371/journal.pone.0222916

Editor: Quanquan Gu, UCLA, UNITED STATES

Received: February 12, 2019

Accepted: September 10, 2019

Published: September 26, 2019

Data Availability Statement: All relevant data are within the paper.

Funding: The authors are supported by Ministerio de Ciencia, Innovación y Universidades del Gobierno de España, project ref. PGC2018 - 097848 - B - I0.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Classification is one of the cornerstones of Supervised Machine Learning. In parallel to the development of different methodologies that allow the construction of classifiers, the evaluation process of the classifiers to compare them, and the choice of the best among those available, has caught the attention of researchers.

Introduction of an adequate performance measure for classifiers is a subject no yet closed up to date (see [1]-[3]), and different metrics have been introduced. Some measures are naturally introduced in the binary case, such as Accuracy, Sensitivity, Specificity and Area Under the ROC Curve (AUC), among others, but not all of them can be well extended to the multi-class setting.

One of the ones that does is Accuracy (i.e. the fraction of well-predicted cases over the total), which seems the most natural measure and has been used for decades. Notwithstanding, Accuracy is not an effective measure since, among other things, it does not take into account the distribution of the misclassification among classes nor the marginal distributions. Other more subtle measures have been introduced in the multi-class setting to address this issue, improving efficiency and class discrimination power.

We will focus our attention in Matthews Correlation Coefficient (MCC) and Cohen's *Kappa*. The former was introduced in the binary setting by Matthews ([4]), and generalized to the multi-class case in [5], being commonly used as a reference performance measure, especially for unbalanced data sets, in different fields as, for example, bioinformatics (see [5]-[7]). On the other hand, *Kappa* is a traditional measure originally designed as a measure of agreement between two judges, based on the Accuracy but corrected for chance agreement. At present, its use is not simply limited to medicine or psychology (see for instance, [8] and [9]), but is a measure widely used in other fields as ecology ([10] and [11]), neuroscience ([12]) or machine learning, where it is used to evaluate the agreement between the actual and the assigned classes by a classifier. In the classification literature, the discussion on *Kappa* is most focused on its suitability compared to other classifiers; for example, in [1] *Kappa* has been considered jointly with 17 other performance metrics in several scenarios.

It is not an overstatement to say that *Kappa* is one of the most widespread measures and of use in several fields and disciplines. Nevertheless, some authors, including the introducer of *Kappa* statistic himself, Jakob Cohen, alerted that *Kappa* could be inadequate in different circumstances, specifically when an imbalance distribution of classes is involved, i.e. the marginal probability of one class is much more (or less) greater than the others (leaving aside the literature below, on which we will deal more closely, see also [13]-[17]). According to them, some problems arise in such situations because it is not clear how the hypothetical probability of chance agreement should be defined. In [18] and [19], the so-called *Kappa paradox* is described. Roughly speaking, *Kappa paradox* arises since for a fixed agreement between judges, the *Kappa* statistic penalizes judges with similar marginals compared with judges with different ones. The authors show several examples where this happens.

This same obstacle is extensively studied in [20]-[22]. In the later, two separate causes of the *paradox* are considered; (1) the *prevalence paradox* arises from the fact that when the hypothetical probability of chance agreement among raters is high, even high values of the relative observed agreement (which is identical to Accuracy) produce low values of *Kappa*, and (2) the *bias paradox*, which is the consequence of the fact that imbalanced marginal distributions produce higher scores of *Kappa*. The authors claim that reporting a single agreement coefficient makes interpretation and comparison difficult. Hence, they suggest a corrected version of *Kappa* for *bias* and *prevalence* (PABAK), which should be used together with *Kappa*.

Similar conclusions emerge from [23], where the authors claim that *Kappa* is a relative measure of agreement, which is an inadequate characteristic for assessing in a clinical setting, specifically if a high agreement among experts leads to lower values of *Kappa*. Instead, they suggest using *the proportion of specific agreement* ([24]), which divides the agreement into a positive and a negative rate, allowing professionals to have an absolute measure and at the same time, information about the marginal distributions. Regarding the effect on estimation of the chance agreement, Albatine et al. ([25]) analysed 28 different similarity measures for clustering purposes; they suggest adding a correction for chance, in a specific family of coefficients, which makes some of them equivalent, regardless of how expectations are calculated. This work is extended by Warrens in [26], where more in-depth analysis is presented and several indices are generalized: Cohen's kappa ([27]), Scott's pi ([28]), Mak's rho([29]), Goodman and Kruskal's lambda ([30]), and Hamann's eta ([31]).

On the other hand, there are several authors that defend that *Kappa* is a useful measure of agreement, when its limitations are taken into account. For example, in [32] the authors defend the use of *Kappa* in a previous study, and warn that it is a useful measure if marginal distributions are considered. A similar conclusion was reached in [33], where it is said that although *Kappa* is not suitable in certain circumstances, it is better than the raw proportion. In [34] the work of [22] expands and the *Kappa* pitfalls are explained for the agreement between judgments, concluding that if it is used and interpreted properly, the *Kappa* coefficient provides a valuable information. As in previous works, they propose to use corrected versions of the coefficient as well. In [16] the author argues that in the case of dichotomous variables, *Kappa* is satisfactory (although it is not for other cases); as we show in the present work, even in the binary case, *Kappa* can exhibit unexpected behaviour. Finally, there are some authors ([34]) who do not agree with the use of weighted versions of the statistics as PABAK, and suggest select the marginal distributions to be similar.

In general, the use of *Kappa* is not only extended but accepted, and its pitfalls are overcome by considering the marginal distributions and using weighted alternatives, as, for example the one suggested by Cohen ([15]), PABAK or other alternatives ([35] and [36]).

Despite the vast amount of existing literature, in the field of medicine and psychology, pointing out the threats of *Kappa*, when Classification Machine Learning methods experimented their boom Cohen's *Kappa* was introduced as a reliable performance metric. Actually it is incorporated in the most extended software packages, such as *SciKit Learn* [37] for Python, and *Caret* [38] for R. What is more, in recent studies such as [39]-[42] and [12], *Kappa* is still used as if it were a reliable performance metric. In fact, the literature reviewed recognizes the difficulty of clinical professionals in interpreting *Kappa* because it is a relative measure, that is, *Kappa* itself is not enough to know if two professionals agree or disagree. This does not seem to be a problem in machine learning classification because the ground-truth is always compared with different methods in the same condition of marginal distributions. Therefore, it can be argued that we are not interested in the value of *Kappa* itself (as are the clinicians), but in the difference of the classifying pairs ground-truth, so *Kappa* is a reliable metric for this task. However, the reality is that this is not always the case. As we show, there are scenarios in which, given the same ground-truth, a better classifier can obtain a lower value of *Kappa*. It is important to mention that some authors also highlight the problems associated with *Kappa* when it is used as a performance metric in classification (see for instance [43]-[45]), although they do not perform an exhaustive analysis like the one presented here.

Clearly, marginal distributions seem to play a key role in the problems surrounding *Kappa*. However, there is a lack of a consistent and satisfactory description of the cases in which the unwanted behaviour of *Kappa* appears, and how this affects its use as a performance metric for classification.

In our paper, we deepen the study of the pitfalls discussed above by analysing in detail the unwanted behaviour of *Kappa* from a novel perspective. Our point of view is the identification of situations in which discrepancies in its behaviour, with respect to that of MCC, become evident, going in the opposite direction. Indeed, we study varied scenarios of misclassification in settings with different marginal probabilities of the categories, and how this scenarios affect the statistics *Kappa* and MCC, by analysing both the asymmetry and the entropy of the confusion matrix. Considering *Kappa* as a relative measure of agreement, we provide a mathematical framework to understand the associated problems with it when dealing with extreme unbalanced marginal distributions, which is frequent in machine learning problems.

Our goal is to present a systematic study, both analytical and by means of empirical experimentation, to compare the two performance measures. For that, we investigate the similarities and differences in the behaviour of MCC and *Kappa* in different scenarios. In some of them,

they are strongly correlated, and we show some mathematical relations and study some limit cases. But in others, they exhibit very different behaviour, being that of *Kappa* contrary to common sense, to the point that we join the detractors of its use for the assessment of classifiers. This paper is an attempt to shed some light on the identification of the latter.

The paper is organized as follows: first, we introduce some definitions and state some notations. Next, we prove that if the confusion matrix, which allows visualization of the performance of a classifier, is symmetric, then *Kappa* and MCC coincide. Each column in the confusion matrix represents the cases in any predicted class, while each row represents the cases in any actual class. In the sequel, we study in some detail the binary case, in which classes are named "positive" and "negative" and the confusion matrix has a general form $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, where a = *true positive*, b = *false negative*, c = *false positive* and d = *true negative*, splitting the study according to whether $c = 0$, the scenario in which *Kappa* has a behaviour consistent with that of MCC, and $c > 0$, in which the opposite happens. For each of these cases, we consider particular sub-cases and we deepen in their study. We also consider a pathological multi-class unbalanced situation, in which one of the classes is much more common than the others, and it is mainly misclassified (family of confusion matrices Z_A introduced in [2]). We also perform empirical experimentation in dimension 3, considering some families of confusion matrices, and finish with a few concluding words.

Definitions and notations

Given a generic matrix M, let M^T denote its transpose, that is, the matrix obtained from M by interchanging columns and rows. The same notation applies to vectors, which by default are column vectors. We say that matrix Q is equivalent to M, and denote it by $Q \equiv M$, if Q can be obtained from M by multiplying it by a positive constant.

Classification

Classification consists of assigning a case to a class (category or label) on the basis of a known set of features or characteristics. This is usually done by a classifier learned from a training dataset. From the validation process of the classifier with a testing dataset, we obtain a confusion matrix C, which takes into account actual and predicted classes of the cases in the testing dataset. To fix ideas, assume that there are N different classes labeled $\{1, \ldots, N\}$. Then, $C = (C_{ij})_{i,j=1,\ldots,N}$ is a $N \times N$ matrix defined by: C_{ij} is the number of cases in the testing dataset that belong to class i and have been assigned to class j by the classifier. Note that $C_{ij} \geq 0$. Let S denote the sum of all the elements of C (the number of cases in the testing dataset), that is, $S = \sum_{i,j=1}^{N} C_{ij} > 0$. In the binary case $N = 2$, to abbreviate notation we preferably denote $C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$ by $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, as previously mentioned in the Introduction.

In the context of classification, *Accuracy* (Acc for brief) is the fraction of correctly classified cases in the testing dataset, that is, $\mathrm{Acc} = \sum_{i=1}^{N} C_{ii}/S$. This performance measure is one of the most intuitive, and it is naturally extended to multi-class from binary classification. Acc mainly considers the diagonal of the confusion matrix, and does not take into account how the off-diagonal elements, corresponding to misclassification, are distributed.

Other more subtle performance measures based on the confusion matrix have been introduced to compare classifiers. We here compare two of the most commonly used. Note that these measures are invariant for equivalent confusion matrices.

Matthews correlation coefficient

The binary case. *Matthews Correlation Coefficient* MCC was first introduced in the binary case by B.W. Matthews [4] to assess the performance of protein secondary structure prediction, as the ϕ-coefficient, which is the measure of association obtained by discretization of the Pearson's correlation coefficient for two binary vectors. That is, in the binary case, MCC = $\phi = \rho(x, y)$, where $x = (x_1, \ldots, x_S)^T$ and $y = (y_1, \ldots, y_S)^T$ are the S-dimensional binary vectors defined in this way:

$$x_i = \begin{cases} 1 & \text{if case } i \text{ belongs to class "positive",} \\ 0 & \text{if it belongs to class "negative",} \end{cases}$$

$$y_i = \begin{cases} 1 & \text{if case } i \text{ has been classified as belonging to class "positive",} \\ 0 & \text{if it has been classified as belonging to class "negative",} \end{cases}$$

and ρ is Pearson's correlation coefficient defined by

$$\rho(x,y) = \frac{Cov(x,\, y)}{\sqrt{Cov(x,\, x)}\,\sqrt{Cov(y,\, y)}} = \frac{\sum_{i=1}^{S}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{S}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{S}(y_i - \bar{y})^2}} \tag{1}$$

where, as usual, $\bar{x} = \frac{1}{S}\sum_{i=1}^{S} x_i$ and $\bar{y} = \frac{1}{S}\sum_{i=1}^{S} y_i$, and $Cov(x, y)$ denotes the statistical covariance of x and y, that is, $Cov(x,\, y) = \frac{1}{S}\sum_{i=1}^{S}(x_i - \bar{x})\,(y_i - \bar{y})$, and when $x = y$, $Cov(x, x) = Var(x)$ is the statistical (uncorrected) variance of x. Note that the square of the ϕ-coefficient is related to the chi-squared statistic for the 2×2 contingency table, χ^2, by means of $\phi^2 = \frac{\chi^2}{S}$. Then, using some algebra and taking into account that, by definition of vectors x and y, the elements of the confusion matrix are

$$a = \sum_{i=1}^{S} x_i\, y_i, \quad b = \sum_{i=1}^{S} x_i\,(1 - y_i), \quad c = \sum_{i=1}^{S}(1 - x_i)\, y_i \;\text{ and }\; d = \sum_{i=1}^{S}(1 - x_i)\,(1 - y_i),$$

we obtain that

$$a\,d - b\,c = S\sum_{i=1}^{S} x_i\, y_i - \Big(\sum_{j=1}^{S} x_j\Big)\Big(\sum_{k=1}^{S} y_k\Big),$$

$$a + b = \sum_{i=1}^{S} x_i, \quad b + d = S - \sum_{i=1}^{S} y_i, \quad a + c = \sum_{i=1}^{S} y_i, \quad c + d = S - \sum_{i=1}^{S} x_i$$

and then using $x_i^2 = x_i$ and $y_i^2 = y_i$ for any $i = 1, \ldots, S$, we can rewrite (1) as

$$\text{MCC} = \frac{a\,d - b\,c}{\sqrt{(a + b)\,(b + d)\,(a + c)\,(c + d)}} \quad \text{(in the binary case).} \tag{2}$$

The multi-class case. In [5] the problem of evaluation of prediction of RNA secondary structure in cases where some predicted pairs go into the category of "unknown" due to lack of reliability, is considered. By introducing an extended correlation coefficient that applies to any number of categories, the author facilitates addressing the problem of predicting base pairs of RNA secondary structure as a three-category problem instead of artificially force it to fall into the binary case by fixing one of the categories, and then considering which cases belong and

which do not belong to that category, leading to a loss of information and a suboptimal procedure. Indeed, MCC is generalized in [5] to classification with $N > 2$ classes based on considering the expected covariance of all categories and constructing the following extension of Pearson's correlation coefficient ρ from a pair of binary vectors to a pair of binary matrices:

$$\tilde{\rho}(X, Y) = \frac{\widetilde{Cov}(X,\ Y)}{\sqrt{\widetilde{Cov}(X,\ X)}\sqrt{\widetilde{Cov}(Y,\ Y)}}\ , \tag{3}$$

where if X and Y are two matrices $S \times N$, $\widetilde{Cov}(X,\ Y)$ is defined as the average of the N covariances between the different pairs of S-dimensional binary vectors given by the same column in matrices X and Y, that is, $\widetilde{Cov}(X,\ Y) = \frac{1}{N}\ \sum_{k=1}^{N} Cov(x^k,\ y^k)$, where $x^k = (X_{1k}, \ldots, X_{Sk})^T$ and $y^k = (Y_{1k}, \ldots, Y_{Sk})^T$ are the columns k of matrices X and Y, respectively. Therefore, by defining $S \times N$ matrices $X = (X_{ij})_{i,j}$ and $Y = (Y_{ij})_{i,j}$ in the following way:

$$X_{ij} = \begin{cases} 1 & \text{if case } i \text{ belongs to class } j, \\ 0 & \text{if it belongs to other class,} \end{cases}$$

$$Y_{ij} = \begin{cases} 1 & \text{if case } i \text{ has been classified as belonging to class } j, \\ 0 & \text{if it has been classified as belonging to other class,} \end{cases}$$

for $i = 1, \ldots, S$ and $j = 1, \ldots, N$, we finally introduce the multi-class extension by $\mathrm{MCC} = \tilde{\rho}(X,\ Y)$, and by using some algebra and that by definition of matrices X and Y, $C_{ij} = \sum_{\ell=1}^{S} X_{\ell i}\, Y_{\ell j}$, we obtain the known expression

$$\mathrm{MCC} = \frac{\sum_{k,\ell,m=1}^{N} (C_{kk}\, C_{\ell m} - C_{mk}\, C_{k\ell})}{\sqrt{\sum_{k=1}^{N}\Big((\sum_{\ell=1}^{N} C_{k\ell})(\sum_{u,v=1,\ u\neq k}^{N} C_{uv})\Big)}\sqrt{\sum_{k=1}^{N}\Big((\sum_{\ell=1}^{N} C_{\ell k})(\sum_{u,v=1,\ u\neq k}^{N} C_{vu})\Big)}} \tag{4}$$

We give below a sketch of the proof of the equivalence between (3) and (4). Indeed, the numerator of (3) can be developed as follows:

$$\begin{aligned} \widetilde{Cov}(X,\ Y) &= \frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{S}\sum_{r=1}^{S}\left(X_{rk} - \bar{x^k}\right)\left(Y_{rk} - \bar{y^k}\right)\right) = \frac{1}{N\,S}\sum_{k=1}^{N}\left(\sum_{r=1}^{S} X_{rk}\, Y_{rk} - S\,\bar{x^k}\,\bar{y^k}\right) \\ &= \frac{1}{N\,S^2}\sum_{k,\ell,m=1}^{N}(C_{kk}\, C_{\ell m} - C_{k\ell}\, C_{mk}) \end{aligned}$$

using that $S\,\bar{x^k}\,\bar{y^k} = \frac{1}{S}\ \sum_{\ell,m=1}^{N} C_{k\ell}\, C_{mk}$, which is a consequence of the fact that by definition, $\bar{x^k} = \frac{1}{S}\sum_{r=1}^{S} X_{rk} = \frac{1}{S}\sum_{\ell=1}^{N} C_{k\ell}$ since

$$\sum_{\ell=1}^{N} C_{k\ell} = \sum_{\ell=1}^{N}(\sum_{r=1}^{S} X_{rk}\, Y_{r\ell}) = \sum_{r=1}^{S} X_{rk}\,(\sum_{\ell=1}^{N} Y_{r\ell}) = \sum_{r=1}^{S} X_{rk}$$

(note that by definition of Y, $\sum_{\ell=1}^{N} Y_{r\ell} = 1$), and analogously with $\bar{y^k} = \frac{1}{S}\sum_{r=1}^{S} Y_{rk} = \frac{1}{S}\sum_{m=1}^{N} C_{mk}$.

We also used that $\sum_{r=1}^{S} X_{rk}\, Y_{rk} = C_{kk}$, and that $S = \sum_{\ell,m=1}^{N} C_{\ell m}$. Now we develop the term in the denominator of (3) corresponding to X (analogous development would be obtained for Y):

$$\begin{aligned}
\widetilde{Cov}(X,\,X) &= \frac{1}{N}\sum_{k=1}^{N}\Big(\frac{1}{S}\sum_{r=1}^{S}\big(X_{rk}-\bar{x^k}\big)^2\Big) = \frac{1}{N\,S}\sum_{k=1}^{N}\left(\sum_{r=1}^{S}X_{rk}^2 - S\,\big(\bar{x^k}\big)^2\right)\\
&= \frac{1}{N\,S}\sum_{k=1}^{N}\Big(\sum_{\ell=1}^{N}C_{k\ell} - \frac{1}{S}\left(\sum_{v=1}^{N}C_{kv}\right)^2\Big) \quad \text{where we use that } X_{rk}^2 = X_{rk}\\
&= \frac{1}{N\,S^2}\sum_{k=1}^{N}\left(\left(\sum_{\ell=1}^{N}C_{k\ell}\right)\left(S-\sum_{v=1}^{N}C_{kv}\right)\right)\\
&= \frac{1}{N\,S^2}\sum_{k=1}^{N}\left(\left(\sum_{\ell=1}^{N}C_{k\ell}\right)\left(\sum_{u,v=1,u\neq k}^{N}C_{uv}\right)\right) \quad \text{using that } S=\sum_{u,v=1}^{N}C_{uv}.
\end{aligned}$$

Note that in the binary case, expression (4) matches (2). Indeed, when $N = 2$, numerator of (4) can be written as $2(C_{11}\,C_{22} - C_{21}\,C_{12}) = 2(ad - bc)$, while the first term in the denominator is $\sqrt{2\,(a+b)\,(c+d)}$, and the second one coincides with $\sqrt{2\,(a+c)\,(b+d)}$.

Software provided by the author of [5] allowing to perform the calculations easily is available at http://rk.kvl.dk/.

Cohen's *Kappa*

Cohen's *Kappa* statistic, or simply *Kappa* (henceforth, also denoted by $\mathcal{K}$), was originally introduced by J. A. Cohen [27] in the field of psychology as a measure of agreement between two judge, and later it has been used in the literature as a performance measure in classification, as for example in [46]. More concretely, *Kappa* is used in classification as a measure of agreement between observed and predicted or inferred classes for cases in a testing dataset. Its definition is:

$$\mathcal{K} = \frac{\mathrm{Acc} - P_e}{1 - P_e}\,, \tag{5}$$

where P_e is the hypothetical probability of chance agreement, using the values of the confusion matrix to estimate the probabilities of randomly choose each class, that is, $P_e = \sum_{i=1}^{N}\frac{C_{i\,\cdot}\times C_{\cdot\,i}}{S^2}$, where as usual, we use the notations $C_{i\,\cdot} = \sum_{j=1}^{N} C_{ij}$ (the sum of row i), and $C_{\cdot\,j} = \sum_{\ell=1}^{N} C_{\ell j}$ (the sum of column j).

Both MCC and *Kappa* assume their theoretical maximum value of +1 when classification is perfect, the larger the metric value, the better the classifier performance. MCC ranges between −1 and +1 while *Kappa* does not in general, although it does in the cases considered in this work. Moreover, it is straightforward to see that they are symmetric, that is, $\mathcal{K}(C^T) = \mathcal{K}(C)$ and $\mathrm{MCC}(C^T) = \mathrm{MCC}(C)$.

The symmetric case

In the case of a symmetric confusion matrix, it is known that *Kappa* statistic is equivalent to *Scott's pi* ([28], [47]), which is a special case of *Krippendorff's alpha* ([48]). *Scott's pi* is a statistic with the same structure as *Kappa* but that differs from it in the definition of P_e. Hereunder, we will show that if C is a symmetric matrix, *Kappa* and MCC not only are consistent with each other but they coincide exactly. Although this result seems to be known, we could not find a reference for it and therefore, we provide its proof here.

Proposition 1 *Let $C = (C_{ij})_{i,j=1,\dots,N}$ be a symmetric confusion matrix in the general multi-class setting. That is, $C = C^T$. Then, $\mathcal{K}(C) = \mathrm{MCC}(C)$.*

Proof. By (4) and taking into account that $C_{ij} = C_{ji}$ by symmetry, we can write

$$\begin{aligned}\mathrm{MCC}(C) &= \frac{\sum_{k,\ell,m=1}^{N}(C_{kk}C_{\ell m} - C_{mk}C_{k\ell})}{\sum_{k=1}^{N}((\sum_{\ell=1}^{N}C_{k\ell})(\sum_{u,v=1,u\neq k}^{N}C_{uv}))} = \frac{\sum_{k=1}^{N}C_{kk}(\sum_{\ell,m=1}^{N}C_{\ell m}) - \sum_{k,\ell,m=1}^{N}C_{km}C_{k\ell}}{\sum_{k=1}^{N}((\sum_{\ell=1}^{N}C_{k\ell})(S - \sum_{v=1}^{N}C_{kv}))} \\ &= \frac{S\sum_{k=1}^{N}C_{kk} - \sum_{k=1}^{N}(\sum_{\ell=1}^{N}C_{k\ell})^2}{S\sum_{k,\ell=1}^{N}C_{k\ell} - \sum_{k=1}^{N}(\sum_{\ell=1}^{N}C_{k\ell})^2} = \frac{S\sum_{k=1}^{N}C_{kk} - \sum_{k=1}^{N}(\sum_{\ell=1}^{N}C_{k\ell})^2}{S^2 - \sum_{k=1}^{N}(\sum_{\ell=1}^{N}C_{k\ell})^2}\,.\end{aligned} \tag{6}$$

On the other hand, by symmetry we can write $P_e = \sum_{k=1}^{N} C_{k\cdot}^2/S^2$, and therefore,

$$\mathcal{K}(C) = \frac{\frac{\sum_{k=1}^{N}C_{kk}}{S} - \frac{\sum_{k=1}^{N}(\sum_{\ell=1}^{N}C_{k\ell})^2}{S^2}}{1 - \frac{\sum_{k=1}^{N}(\sum_{\ell=1}^{N}C_{k\ell})^2}{S^2}} = \frac{S\sum_{k=1}^{N}C_{kk} - \sum_{k=1}^{N}(\sum_{\ell=1}^{N}C_{k\ell})^2}{S^2 - \sum_{k=1}^{N}(\sum_{\ell=1}^{N}C_{k\ell})^2},$$

which coincides with MCC(*C*) by (6).

The binary case

Let C be a generic confusion matrix in dimension 2, $C = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. By (2) and (5), we have that

$$\begin{aligned}\mathrm{MCC}(C) &= \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+c)(c+d)}} \quad \text{and} \\ \mathcal{K}(C) &= \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)}\end{aligned}$$

and it turns out that $\mathcal{K}(C)$ is the harmonic mean of α and β, while MCC(*C*) is their geometric mean, being

$$\alpha = \frac{ad - bc}{(a+b)(b+d)} \quad \text{and} \quad \beta = \frac{ad - bc}{(a+c)(c+d)}\,.$$

That is, $\mathcal{K}(C) = \frac{2}{\frac{1}{\alpha}+\frac{1}{\beta}}$ and $\mathrm{MCC}(C) = \sqrt{\alpha\beta}$. As a direct consequence of the known relationship between these two means, we have that in the binary case:

$$\begin{gathered}\min(\alpha, \beta) \leq \mathcal{K}(C), \ \mathrm{MCC}(C) \leq \max(\alpha, \beta) \quad \text{and} \\ \begin{cases} \text{If } ad > bc, & 0 < \mathcal{K}(C) \leq \mathrm{MCC}(C), \\ \text{If } ad < bc, & \mathrm{MCC}(C) \leq \mathcal{K}(C) < 0, \\ \text{If } ad = bc, & \mathrm{MCC}(C) = \mathcal{K}(C) = 0. \end{cases}\end{gathered} \tag{7}$$

Now we delve a little deeper into the relationship between the two performance measures. By the property of invariance for equivalent confusion matrices, we can split the study of the binary case into two different scenarios: $c = 0$ and $c = 1$ (the latter corresponding to $c > 0$). These two cases cover all the possibilities, determining a partition of the set of binary confusion matrices into two subsets with clearly differentiated behaviour. As we will see next, when $c = 0$ there is an agreement between MCC and *Kappa*. What is more, MCC and *Kappa* are linked by means of a functional relationship (see Proposition 2 below) that easily shows the relationship of monotony between them, which implies that when one of them grows or decreases, the other also does the same, that is, they have a consistent behaviour. On the contrary, when $c = 1$ an important disagreement between the two measures highlights in different particular scenarios (see Corollaries 4, 5 and 6). Indeed, in all of them it is shown that while MCC monotonically decreases as the task done by the classifier is getting worse, *Kappa* does not.

Moreover, as the row sums are the actual number of cases in the testing dataset belonging to each class, we assume that they are both strictly positive, that is, $a + b > 0$ and $c + d > 0$. We also must ensure that MCC can be calculated, i.e, that we do not divide by zero. For that, the sum of the columns must also be strictly positive, that is, we additionally assume that $a + c > 0$ and $b + d > 0$.

The c = 0 case: Agreement between MCC *and Kappa*

This case corresponds to perfect classification of the negative class, since there are no cases of the negative class in the testing dataset that have been classified as belonging to the positive class. Then, we assume $a > 0$ and $d > 0$. Moreover, we assume $b > 0$ since $b = 0$ corresponds to the symmetric case already studied in the previous section, in which $\mathcal{K} = \mathrm{MCC} = 1$. We use notation $C_0 = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$. We have, then,

$$\mathrm{MCC}(C_0) = \sqrt{\frac{a\,d}{(a+b)(b+d)}} \quad \text{and} \quad \mathcal{K}(C_0) = \frac{2\,a\,d}{a\,d + (a+b)\,(b+d)}\,.$$

We will show that in this case there is agreement between the behaviour of the two measures. Indeed, they are linked by means of a functional relationship, as can be seen in the next proposition.

Proposition 2

$$\mathcal{K}(C_0) = \frac{2\,(\mathrm{MCC}(C_0))^2}{1 + (\mathrm{MCC}(C_0))^2}\,,$$

and the following properties hold:

1. *Since* $\mathrm{MCC}(C_0) > 0$, $\mathcal{K}(C_0)$ *is a monotonically increasing function of* $\mathrm{MCC}(C_0)$, *so they are consistent performance measures.*
2. $0 < \frac{ad}{(a+b)\,(b+d)} < \mathcal{K}(C_0) < \mathrm{MCC}(C_0) < 1$.
3. *The maximum distance between them is achieved when* $\mathrm{MCC}(C_0) \approx 0.3$, *and is* ≈ 0.13.

Moreover,

- Fixed a, d,

$$\lim_{b\to+\infty} \mathrm{MCC}(C_0) = \lim_{b\to+\infty} \mathcal{K}(C_0) = 0\,,$$

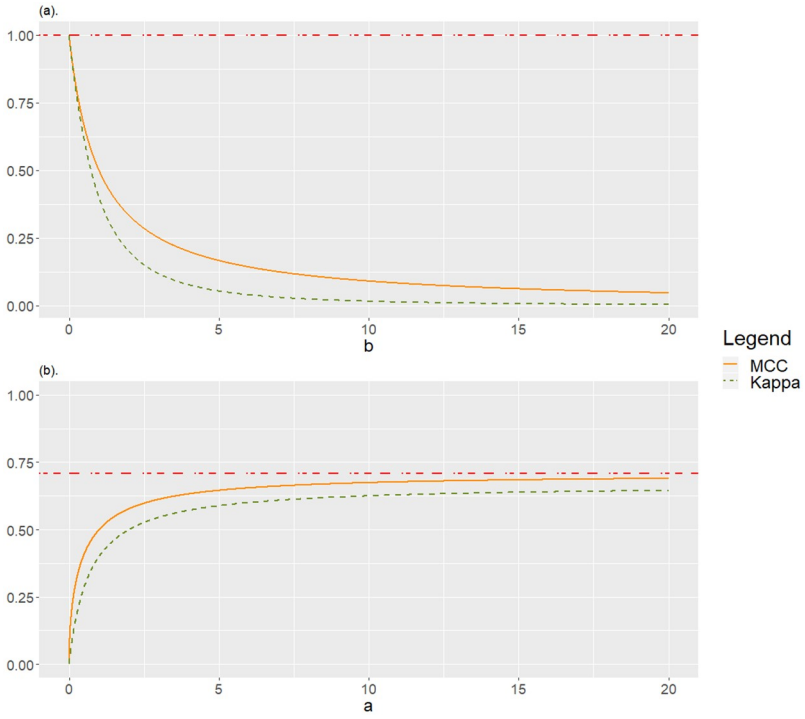

Fig 1. Agreement between MCC and *Kappa* for C_0. Unbalanced case with underrepresentation of the negative class, which is perfectly classified. (a) With $a = d = 1$, as function of b: positive class mainly misclassified. (b) With $b = d = 1$ as function of a: positive class mainly well classified.

https://doi.org/10.1371/journal.pone.0222916.g001

which corresponds to an scenario in which the negative class is underrepresented and cases actually in the positive class are mainly misclassified. On the other hand,

$$\lim_{b\to 0} \mathrm{MCC}(C_0) = \lim_{b\to 0} \mathcal{K}(C_0) = 1\,,$$

corresponding to perfect classification (see Fig 1(a)).

- Fixed b, d,

$$0 < \lim_{a\to+\infty} \mathcal{K}(C_0) = \frac{2\,d}{2\,d+b} < \lim_{a\to+\infty} \mathrm{MCC}(C_0) = \sqrt{\frac{d}{b+d}} < 1\,,$$

which corresponds to an scenario in which the negative class is underrepresented but cases actually in the positive class are mainly well classified. Note that as $b \to 0$, both $\lim_{a\to+\infty} \mathcal{K}(C_0)$ and $\lim_{a\to+\infty} \mathrm{MCC}(C_0)$, tend to be 1.
On the other hand,

$$\lim_{a\to 0} \mathrm{MCC}(C_0) = \lim_{a\to 0} \mathcal{K}(C_0) = 0\,,$$

corresponding to complete misclassification of the positive class (see Fig 1(b)).

- The case with a, b fixed, considering $\mathrm{MCC}(C_0)$ and $\mathcal{K}(C_0)$ as function of d, is symmetric to the previous one, and then omitted.

The c = 1 case: Disagreement between MCC *and Kappa*

This case corresponds to not-completely perfect classification of the negative class, since there is at least one case in the testing dataset belonging to this class that has been classified as being in the positive class. We assume $b > 0$ since if $b = 0$ we are in the previous situation, by symmetry of MCC and *Kappa*. Although $b = 1$ corresponds to a symmetric confusion matrix already studied, we include it in this section for the sake of completeness. We use the notation $C_1 = \begin{pmatrix} a & b \\ 1 & d \end{pmatrix}$. Then,

$$\begin{aligned} \mathrm{MCC}(C_1) &= \frac{a\,d - b}{\sqrt{(a+1)\,(a+b)\,(d+1)\,(d+b)}}\,, \\ \mathcal{K}(C_1) &= \frac{2\,(a\,d - b)}{(a+1)\,(d+1) + (a+b)\,(d+b)} \end{aligned}$$

Proposition 3 *If* $a = d = 0,\ b \neq 1, \quad -1 = \mathrm{MCC}(C_1) < \mathcal{K}(C_1) = \frac{-2\,b}{1+b^2} < 0\,.$
If $a = d = 0,\ b = 1, \quad \mathrm{MCC}(C_1) = \mathcal{K}(C_1) = -1\,.$
Otherwise,

$$\begin{cases} \textit{if}\ \ ad = b, & \mathrm{MCC}(C_1) = \mathcal{K}(C_1) = 0\,, \\ \textit{if}\ \ ad > b, & \begin{cases} \textit{if}\ \ b > 1, & 0 < \frac{ad-b}{(a+b)(d+b)} < \mathcal{K}(C_1) < \mathrm{MCC}(C_1) < \frac{ad-b}{(a+1)(d+1)} < 1, \\ \textit{if}\ \ b = 1, & 0 < \mathcal{K}(C_1) = \mathrm{MCC}(C_1) = \frac{ad-1}{(a+1)(d+1)} < 1, \\ \textit{if}\ \ b < 1, & 0 < \frac{ad-b}{(a+1)(d+1)} < \mathcal{K}(C_1) < \mathrm{MCC}(C_1) < \frac{ad-b}{(a+b)(d+b)} < 1, \end{cases} \\ \textit{if}\ \ ad < b, & \begin{cases} \textit{if}\ \ b > 1, & \max\left(-1, \frac{ad-b}{(a+1)(d+1)}\right) < \mathrm{MCC}(C_1) < \mathcal{K}(C_1) < \frac{ad-b}{(a+b)(d+b)} < 0, \\ \textit{if}\ \ b = 1, & -1 < \mathrm{MCC}(C_1) = \mathcal{K}(C_1) = \frac{ad-1}{(a+1)(d+1)} < 0, \\ \textit{if}\ \ b < 1, & \max\left(-1, \frac{ad-b}{(a+b)(d+b)}\right) < \mathrm{MCC}(C_1) < \mathcal{K}(C_1) < \frac{ad-b}{(a+1)(d+1)} < 0. \end{cases} \end{cases}$$

Next we consider some particular scenarios of this case that should be explored.

1. $a = d > 0$.

   We use notation $C_{1,a}^{a,b} = \begin{pmatrix} a & b \\ 1 & a \end{pmatrix}$. Fixed $a > 0$, if $b > 1$, the negative class is underrepresented, and the positive class is mainly misclassified, while if $b < 1$, say $b = 1/h$ with $h > 1$, $C_{1,a}^{a,b} \equiv \begin{pmatrix} h\,a & 1 \\ h & h\,a \end{pmatrix}$, which is a confusion matrix that corresponds to underrepresentation of the positive class while it is mainly well classified (if $b \to 0$, which is equivalent to $h \to +\infty$). Then,

$$\mathrm{MCC}(C_{1,a}^{a,b}) = \frac{a^2 - b}{(a+1)\,(a+b)} \quad \text{and} \quad \mathcal{K}(C_{1,a}^{a,b}) = \frac{2\,(a^2 - b)}{(a+1)^2 + (a+b)^2}\,.$$

   From these expressions and Proposition 3, we obtain:
   Corollary 4 *If* $a = d = 0,\ b \neq 1, \quad -1 = \mathrm{MCC}(C_1) < \mathcal{K}(C_1) = \frac{-2\,b}{1+b^2} < 0\,.$
   If $a = d = 0,\ b = 1, \quad \mathrm{MCC}(C_1) = \mathcal{K}(C_1) = -1\,.$

Otherwise,

$$
\begin{cases}
\textit{if } \; a^2 = b, & \mathrm{MCC}(C_1) = \mathcal{K}(C_1) = 0\,, \\
\textit{if } \; a^2 > b, & \begin{cases}
\textit{if } \; b > 1, & 0 < \frac{a^2-b}{(a+b)^2} < \mathcal{K}(C_1) < \mathrm{MCC}(C_1) < \frac{a^2-b}{(a+1)^2} < 1, \\
\textit{if } \; b = 1, & 0 < \mathcal{K}(C_1) = \mathrm{MCC}(C_1) = \frac{a-1}{(a+1)} < 1, \\
\textit{if } \; b < 1, & 0 < \frac{a^2-b}{(a+1)^2} < \mathcal{K}(C_1) < \mathrm{MCC}(C_1) < \frac{a^2-b}{(a+b)^2} < 1,
\end{cases} \\
\textit{if } \; a^2 < b, & \begin{cases}
\textit{if } \; 1 < b < (a+1)^2 + a^2, & \\
& -1 < \frac{a^2-b}{(a+1)^2} < \mathrm{MCC}(C_1) < \mathcal{K}(C_1) < \frac{a^2-b}{(a+b)^2} < 0, \\
\textit{if } \; (a+1)^2 + a^2 \leq b, & -1 < \mathrm{MCC}(C_1) < \mathcal{K}(C_1) < \frac{a^2-b}{(a+b)^2} < 0, \\
\textit{if } \; b = 1, & -1 < \mathrm{MCC}(C_1) = \mathcal{K}(C_1) = \frac{a-1}{(a+1)} < 0, \\
\textit{if } \; \left(b = 1 - \frac{\sqrt{2}}{2} \text{ and } a = \frac{\sqrt{2}-1}{2}\right) \text{ or } \left(b \in \{b_1, b_2\} \text{ and } 0 < a < \frac{\sqrt{2}-1}{2}\right), & \\
& -1 = \frac{a^2-b}{(a+b)^2} < \mathrm{MCC}(C_1) < \mathcal{K}(C_1) < \frac{a^2-b}{(a+1)^2} < 0, \\
\textit{if } \; b_1 < b < b_2 \text{ and } 0 < a < \frac{\sqrt{2}-1}{2}, & \\
& -1 < \mathrm{MCC}(C_1) < \mathcal{K}(C_1) < \frac{a^2-b}{(a+1)^2} < 0, \\
\textit{Otherwise}, & \\
& -1 < \frac{a^2-b}{(a+b)^2} < \mathrm{MCC}(C_1) < \mathcal{K}(C_1) < \frac{a^2-b}{(a+1)^2} < 0,
\end{cases}
\end{cases}
$$

where

$$0 < b_1 = \frac{(1-2\,a) - \sqrt{(1-2\,a)^2 - 8\,a^2}}{2} < b_2 = \frac{(1-2\,a) + \sqrt{(1-2\,a)^2 - 8\,a^2}}{2} < 1\,.$$

Fixed $a > 0$, $-1 < \lim_{b\to+\infty} \mathrm{MCC}(C_{1,a}^{a,b}) = -\dfrac{1}{a+1} < \lim_{b\to+\infty} \mathcal{K}(C_{1,a}^{a,b}) = 0$, *while*

$$0 < \frac{a^2}{(a+1)^2} < \lim_{b\to 0} \mathcal{K}(C_{1,a}^{a,b}) = \frac{2\,a^2}{(a+1)^2 + a^2} < \lim_{b\to 0} \mathrm{MCC}(C_{1,a}^{a,b}) = \frac{a}{a+1} < 1$$

and $\mathrm{MCC}(C_{1,a}^{a,b})$, *as a function of b, is monotonically decreasing when b increases, which agrees with the intuition, since when b monotonically increases, the task done by the classifier is clearly getting worse, while* $\mathcal{K}(C_{1,a}^{a,b})$ *is not. Indeed, fixed* $a > 0$, $\mathcal{K}(C_{1,a}^{a,b})$ *has a global minimum at* $b = b_0$ *with*

$$b_0 = a^2 + (a+1)\sqrt{a^2+1} > a^2\,.$$

See Fig 2 to observe the behaviour of MCC and *Kappa* fixed $a = 0.2$, as function of b.

Remark 1 *Corollary 4 explains the behaviour of* MCC *and Kappa for a confusion matrix equivalent to* $C_{1,a}^{a,b} = \begin{pmatrix} a & b \\ 1 & a \end{pmatrix}$, *according to the values of a = "true positive" = "true negative", and b = "false negative"/"false positive". In particular, fixed "true positive" = "true negative" and "false positive", we observe a contradictory behaviour between these two performance measures as b increases. Indeed, as "false negative"/"false positive" is increasing (implying that the negative class is underrepresented, and the positive class is mainly misclassified),* MCC *monotonically decreases, what is reasonable, but Kappa does not. In fact, Kappa decreases for low values of b* ($b < b_0$) *but increases otherwise. This unreasonable behaviour of Kappa goes in the direction of the thesis defended in this work.* Fig 2 *graphically shows this fact for the particular case a* = 0.2, *corresponding to a confusion matrix equivalent to* $\begin{pmatrix} 1 & 5\,b \\ 5 & 1 \end{pmatrix}$.

Case $b > 1$, with $a = 1$, corresponds to matrix Z_A with $A = b$ and dimension $N = 2$, which is a pathological situation that will be studied in the next section.

2. $a > 0$, $d = 0$.

   We use notation $C_{1,0}^{a,b} = \begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix}$. In this case,

$$\mathrm{MCC}(C_{1,0}^{a,b}) = -\sqrt{\frac{b}{(a+1)\,(a+b)}} \quad \text{and} \quad \mathcal{K}(C_{1,0}^{a,b}) = -\frac{2\,b}{(a+1)+b\,(a+b)}.$$

   and application of Proposition 3 allows obtaining the following result:

   Corollary 5

$$\begin{cases} \text{if } 1 < b < a+1, & -1 < \frac{-b}{a+1} < \mathrm{MCC}(C_1) < \mathcal{K}(C_1) < \frac{-1}{a+b} < 0, \\ \text{if } a+1 \le b, & -1 < \mathrm{MCC}(C_1) < \mathcal{K}(C_1) < \frac{-1}{a+b} < 0, \\ \text{if } b = 1 & -1 < \mathrm{MCC}(C_1) = \mathcal{K}(C_1) = \frac{-1}{a+1} < 0, \\ \text{if } b < 1 < a+b, & -1 < \frac{-1}{a+b} < \mathrm{MCC}(C_1) < \mathcal{K}(C_1) < \frac{-b}{a+1} < 0, \\ \text{if } a+b \le 1, & -1 < \mathrm{MCC}(C_1) < \mathcal{K}(C_1) < \frac{-b}{a+1} < 0. \end{cases}$$

   Although fixed $a > 0$, $\mathrm{MCC}(C_{1,0}^{a,b})$ *is a monotonically decreasing function of b, coinciding with intuition,* $\mathcal{K}(C_{1,0}^{a,b})$ *is not, achieving its global minimum when* $b = \sqrt{a+1}$. *Moreover, fixed* $a > 0$,

$$-1 < \lim_{b\to+\infty} \mathrm{MCC}(C_{1,0}^{a,b}) = -\frac{1}{\sqrt{a+1}} < \lim_{b\to+\infty} \mathcal{K}(C_{1,0}^{a,b}) = 0\,,$$
$$\lim_{b\to 0} \mathrm{MCC}(C_{1,0}^{a,b}) = \lim_{b\to 0} \mathcal{K}(C_{1,0}^{a,b}) = 0\,.$$

   See Fig 3 to observe the behaviour of MCC and *Kappa*, fixed $a = 1$, as function of b.

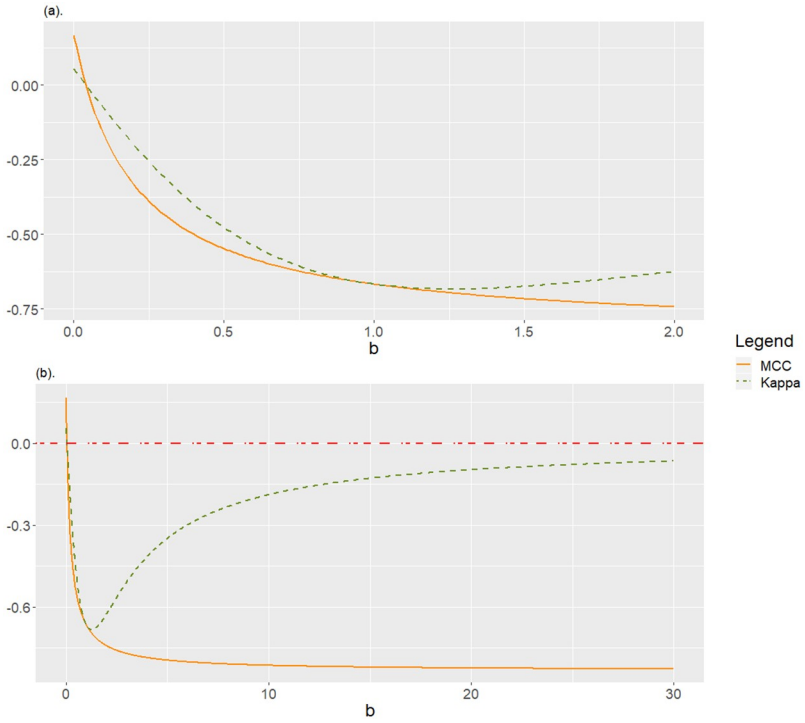

Fig 2. Disagreement between MCC and *Kappa* for $C_{1,a}^{a,b}$ with $a = 0.2$, as function of $b \geq 0$. If $b > 1$, the negative class is underrepresented and quite misclassified, and the positive class is mainly misclassified. (a) A zoom of the detail for $b \leq 2$. (b) For $b \leq 30$.

https://doi.org/10.1371/journal.pone.0222916.g002

Remark 2 *In Corollary 5 we can observe the behaviour of* MCC *and Kappa for a confusion matrix equivalent to* $C_{1,0}^{a,b} = \begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix}$, *corresponding to a scenario in which the negative class is underrepresented and the classifier systematically misclassifies this class, and generally also misclassifies the positive class if b = "false negative"/"false positive" is big. In particular, fixed "true positive" and "false positive", we observe a contradictory behaviour between* MCC *and Kappa as b increases: while* MCC *monotonically decreases, what is expected, Kappa decreases for* $b < \sqrt{a+1}$ *but increases otherwise. Again, we observe here an unreasonable behaviour of Kappa, which is graphically showed in Fig 3 for the particular case a = 1, corresponding to a confusion matrix equivalent to* $\begin{pmatrix} 1 & b \\ 1 & 0 \end{pmatrix}$.

3. $d = 1$, $a \geq 0$.

   We use notation $C_{1,1}^{a,b} = \begin{pmatrix} a & b \\ 1 & 1 \end{pmatrix}$. Classification of negative class is entirely done by random, that is, with the same probability a case actually in the negative class is classified as belonging to any of the two classes. If $a, b > 1$, negative class is underrepresented. We have that

$$\mathrm{MCC}(C_{1,1}^{a,b}) = \frac{a-b}{\sqrt{2\,(a+1)\,(b+1)\,(a+b)}},\ \mathcal{K}(C_{1,a}^{a,b}) = \frac{2\,(a-b)}{2\,(a+1)+(b+1)\,(a+b)}$$

   and application of Proposition 3 gives:

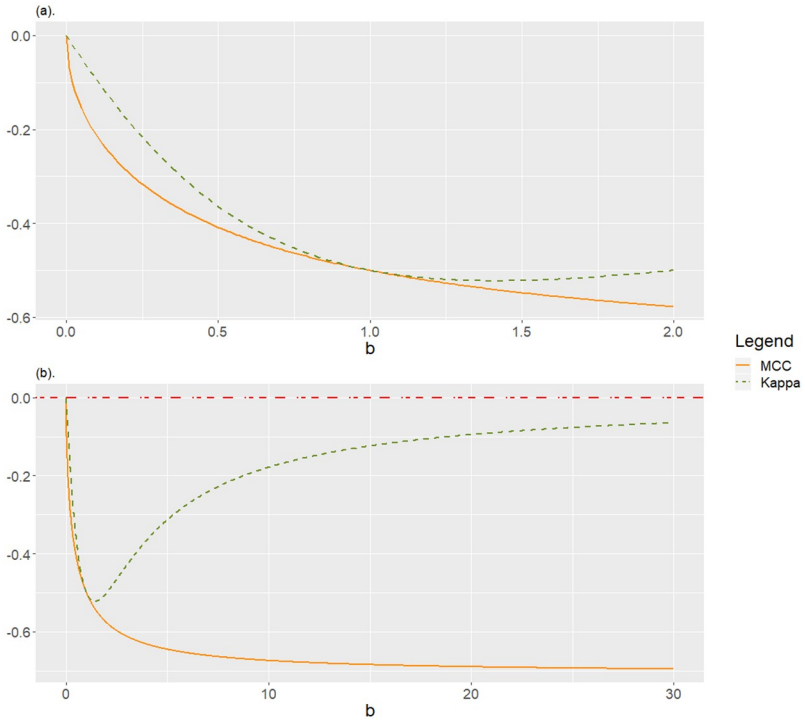

Fig 3. Disagreement between MCC and *Kappa* for $C^{a,b}_{1,0}$ with $a = 1$, as function of $b \geq 0$. If $b > 1$, the negative class is underrepresented and systematically misclassified, and the positive class is also mainly misclassified. (a) A zoom of the detail for $b \leq 2$. (b) For $b \leq 30$.

https://doi.org/10.1371/journal.pone.0222916.g003

Corollary 6

$$
\begin{cases}
\textit{if}\ \ a = b, & \mathrm{MCC}(C^{a,b}_{1,1}) = \mathcal{K}(C^{a,b}_{1,1}) = 0\,, \\
\textit{if}\ \ a > b, & \begin{cases}
\textit{if}\ \ b > 1, & 0 < \frac{a-b}{(a+b)(b+1)} < \mathcal{K}(C^{a,b}_{1,1}) < \mathrm{MCC}(C^{a,b}_{1,1}) < \frac{a-b}{2(a+1)} < 1, \\
\textit{if}\ \ b = 1, & 0 < \mathcal{K}(C^{a,b}_{1,1}) = \mathrm{MCC}(C^{a,b}_{1,1}) = \frac{a-1}{2(a+1)} < 1, \\
\textit{if}\ \ b < 1, & 0 < \frac{a-b}{2(a+1)} < \mathcal{K}(C^{a,b}_{1,1}) < \mathrm{MCC}(C^{a,b}_{1,1}) < \frac{a-b}{(a+b)(b+1)} < 1,
\end{cases} \\
\textit{if}\ \ a < b, & \begin{cases}
\textit{if}\ \ 1 < b < 3a + 2, & \\
-1 < \frac{a-b}{2(a+1)} < \mathrm{MCC}(C^{a,b}_{1,1}) < \mathcal{K}(C^{a,b}_{1,1}) < \frac{a-b}{(a+b)(b+1)} < 0, & \\
\textit{if}\ \ 3a + 2 \leq b, & \\
-1 < \mathrm{MCC}(C^{a,b}_{1,1}) < \mathcal{K}(C^{a,b}_{1,1}) < \frac{a-b}{(a+b)(b+1)} < 0, & \\
\textit{if}\ \ b = 1, & \\
-1 < \mathrm{MCC}(C^{a,b}_{1,1}) = \mathcal{K}(C^{a,b}_{1,1}) = \frac{a-1}{2(a+1)} < 0, & \\
\textit{if}\ \ b < 1, & \\
-1 < \frac{a-b}{(a+b)(b+1)} < \mathrm{MCC}(C^{a,b}_{1,1}) < \mathcal{K}(C^{a,b}_{1,1}) < \frac{a-b}{2(a+1)} < 0. &
\end{cases}
\end{cases}
$$

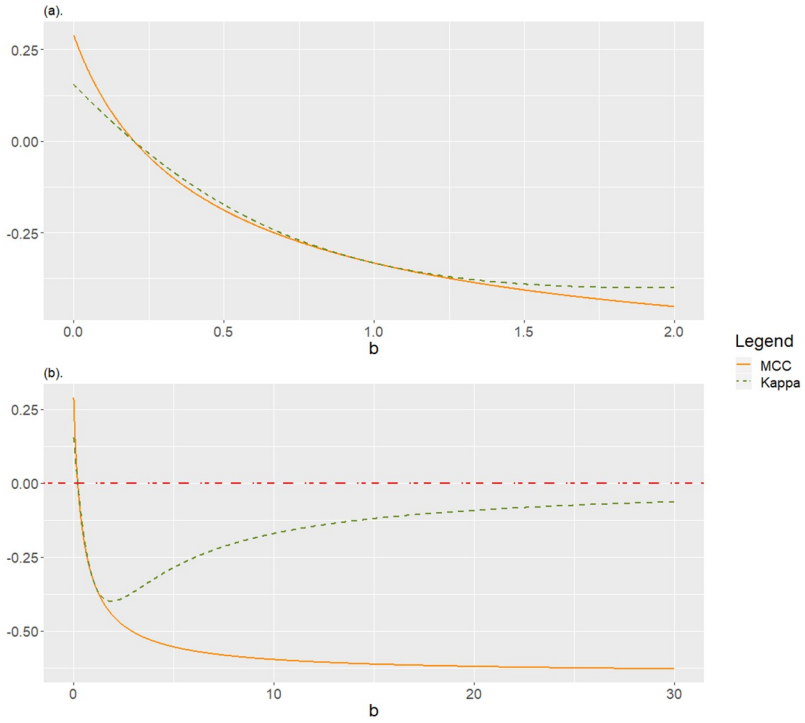

Fig 4. Disagreement between MCC and *Kappa* for $C_{1,1}^{a,b}$ with $a = 0.2$, as function of $b \geq 0$. The negative class is classified at random. If $b > 1$ the positive class is mainly misclassified, and the negative class is underrepresented. (a) A zoom of the detail for $b \leq 2$. (b) For $b \leq 30$.

https://doi.org/10.1371/journal.pone.0222916.g004

As in the previous cases with $c = 1$, although if we fix $a > 0$, then $\mathrm{MCC}(C_{1,1}^{a,b})$ is a monotonically decreasing function of b, coinciding with intuition, we can see that $\mathcal{K}(C_{1,1}^{a,b})$ is not, achieving its global minimum when $b = a + \sqrt{2}\,(a+1)$. Moreover, fixed $a > 0$,

$$-1 < \lim_{b\to+\infty} \mathrm{MCC}(C_{1,1}^{a,b}) = -\frac{1}{\sqrt{2\,(a+1)}} < \lim_{b\to+\infty} \mathcal{K}(C_{1,1}^{a,b}) = 0\,,$$

$$0 < \frac{a}{2\,(a+1)} < \lim_{b\to 0} \mathcal{K}(C_{1,1}^{a,b}) = \frac{2\,a}{3\,a+2} < \lim_{b\to 0} \mathrm{MCC}(C_{1,1}^{a,b}) = \sqrt{\frac{a}{2\,(a+1)}} < 1\,.$$

In Fig 4 we can observe the behaviour of MCC and *Kappa*, fixed $a = 0.2$, as function of b.

Remark 3 *Finally, Corollary 6 is dedicated to confusion matrices equivalent to $C_{1,1}^{a,b} = \begin{pmatrix} a & b \\ 1 & 1 \end{pmatrix}$, which correspond to an unbalanced database set if a, $b > 1$, with minority class the negative one, which is randomly classified, that is, each class is imputed with the same probability to a case actually in the negative class. In addition, if fixed a = "true positive"/"true negative", when b = "false negative"/"false positive" increases the positive class is mainly misclassified. While* MCC *in this situation behaves as expected and monotonically decreases, Kappa does not, increasing for $b > a + \sqrt{2}\,(a+1)$. As in the previous corollaries, an unreasonable behaviour of Kappa is observed, which is shown in Fig 4 for the particular case $a = 0.2$, that is, for a confusion matrix equivalent to $\begin{pmatrix} 1 & 5\,b \\ 5 & 5 \end{pmatrix}$.*

The Z_A family

Finally, we consider another situation that highlights the incoherent behaviour of *Kappa*. $\{Z_A, A \geq 0\}$ has been introduced in [2] as a family of confusion matrices useful to analyse performance measures in unbalanced situations. The definition of Z_A is as follows:

$Z_A = \begin{pmatrix} 1 & 1 & \dots & A \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}$. We denote by $\mathrm{MCC}(A)$ and $\mathcal{K}(A)$, respectively, the MCC and *Kappa* values of matrix Z_A. Note that when $N = 2$, this family is a particular case of iii) with $a = 1$ and $b = A$. Then, we obtain from Corollary 6 the following result:

Corollary 7 *If* $N = 2$,

$$\mathrm{MCC}(A) = \frac{1-A}{2\,(1+A)} \quad \text{and} \quad \mathcal{K}(A) = \frac{2\,(1-A)}{4+(1+A)^2}\,.$$

We have that

$$\begin{cases} \textit{If } A = 1, & \mathcal{K}(A) = \mathrm{MCC}(A) = 0\,, \\ \textit{If } A < 1, & 0 \ < \frac{1-A}{4} < \mathcal{K}(A) < \mathrm{MCC}(A) \ < \frac{1-A}{(1+A)^2} < 1\,, \\ \textit{If } 1 < A < 5, & -1 < \frac{1-A}{4} < \ \mathrm{MCC}(A) < \mathcal{K}(A) \ < \frac{1-A}{(1+A)^2} < 0\,, \\ \textit{If } 5 \leq A, & -1 < \ \mathrm{MCC}(A) < \mathcal{K}(A) \ < \frac{1-A}{(1+A)^2} < 0\,. \end{cases}$$

Although $\mathrm{MCC}(A)$ *is a monotonically decreasing function of A, coinciding with intuition,* $\mathcal{K}(A)$ *is not, achieving its global minimum when* $A = 1 + 2\sqrt{2} > 1$. *Moreover,*

$$\begin{aligned} -1 < \lim_{A\to+\infty} \mathrm{MCC}(A) \ &= -\tfrac{1}{2} < \lim_{A\to+\infty} \mathcal{K}(A) = 0\,, \\ 0 < \tfrac{1}{4} < \lim_{A\to 0} \mathcal{K}(A) \ &= \tfrac{2}{5} < \lim_{A\to 0} \mathrm{MCC}(A) = \frac{1}{2} < 1\,. \end{aligned}$$

We generalize the previous result to any $N \geq 2$ in the following proposition:

Proposition 8

$$\begin{aligned} \mathrm{MCC}(A) \ &= \tfrac{1-A}{(N-1)\,(N^2-2(1-A))}\,, \\ \mathcal{K}(A) \ &= N\,\tfrac{1-A}{(1-A)^2-2\,N\,(N-1)\,(1-A)+N^3\,(N-1)}\,, \end{aligned}$$

and the following properties hold:

1. $\mathrm{MCC}(1) = \mathcal{K}(1) = 0$,
2. $\frac{1}{\mathrm{MCC}(A)} - \frac{1}{\mathcal{K}(A)} = \frac{A-1}{N}$ *and then,*

$$\begin{cases} \textit{If } A < 1, & 0 < \mathcal{K}(A) < \mathrm{MCC}(A) < 1\,, \\ \textit{If } 1 < A, & -1 < \mathrm{MCC}(A) < \mathcal{K}(A) < 0\,, \end{cases}$$

3. $-1 < \lim_{A\to\infty}\mathrm{MCC}(A) = \frac{-1}{2\,(N-1)} < \lim_{A\to\infty}\mathcal{K}(A) = 0\,,$
4. $0 < \lim_{A\to 0}\mathcal{K}(A) = \frac{N}{1+N\,(N-1)\,(N^2-2)} < \lim_{A\to 0}\mathrm{MCC}(A) = \frac{1}{(N-1)\,(N^2-2)} < 1\,,$

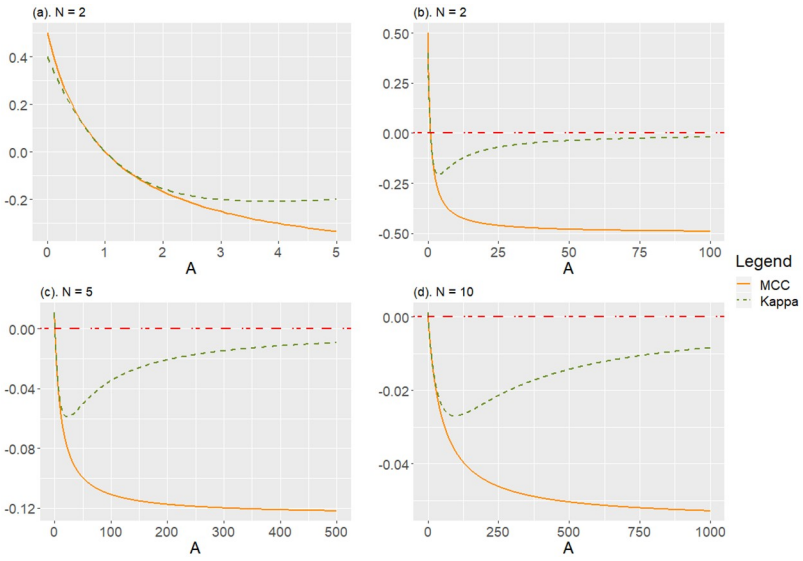

Fig 5. Disagreement between MCC and *Kappa* for Z_A, for different values of *N*. (a) $N = 2$, a zoom of the detail for $A \leq 5$. (b) $N = 2$, $A \leq 100$. (c) $N = 5$, $A \leq 500$. (d) $N = 10$, $A \leq 1000$.

https://doi.org/10.1371/journal.pone.0222916.g005

5. MCC(A) *is monotonically decreasing, while* $\mathcal{K}(A)$ *is not. Indeed,* $\mathcal{K}(A)$ *is a convex function of A, achieving the global minimum, which is a negative value, when* $A = 1 + N\sqrt{N(N-1)}$.
6. *The divergence between* MCC(A) *and* $\mathcal{K}(A)$ *increases monotonically as* $A \to \infty$.

Fig 5 shows the behaviour of MCC and *Kappa* as functions of A, in cases $N = 2$ (both for $A \leq 5$ and for $A \leq 100$), and for $N = 5$ and $N = 10$. A desirable property of any measure of performance is its internal coherence, which implies that if the classifier moves gradually towards a worsening of the classification process, as is the case when A increases for the family Z_A, the measure must reflect this fact with the consequent monotonous decrease (or increase, depending on the interpretation of the measure). Fig 5 highlights the incoherent behaviour of *Kappa*, since as we monotonically increase A, it does not exhibits a monotonic decreasing (as MCC does), and this anomaly not only happens in the binary case ($N = 2$), but continues to occur when we increase N above 2, although at a different scale. Therefore, we have seen that MCC shows internal coherence, unlike *Kappa*, which after decreasing in accordance with the worsening of the classification by increasing A, shows a monotonic growth that goes just in the opposite direction by continuing to increase A, which is clearly inconsistent.

Experimental results

If we recapitulate, we have seen that both in the binary case with $c = 1$, and with the multidimensional Z_A family, as the asymmetry of the confusion matrix increased ($b \to +\infty$ and $A \to +\infty$, respectively), while its diagonal stays constant, the behaviour of *Kappa* and MCC differed more and more. This would be in line with the proven fact that if there is perfect symmetry, therefore these measures match (Proposition 1). It seems natural to ask if it is only the asymmetry that plays a determining role in the discrepancy observed in their linked behaviour (it seems that it should not be like that, since asymmetry of matrix C_0 also increases as $b \to +\infty$, and yet the behaviour of *Kappa* and MCC agree). Or, on the contrary, there is any other characteristic of the matrix that drives in this circumstance. To try to shed some light on this issue, we have carried out some empirical experimentation in dimension $N = 3$.

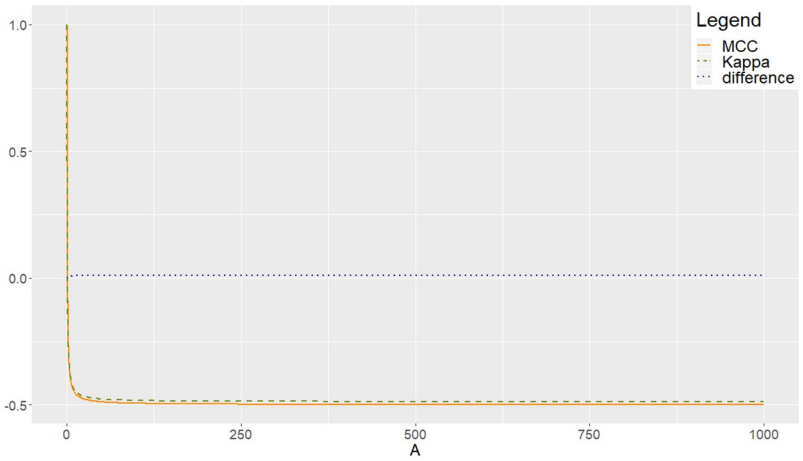

Fig 6. Experimental agreement between MCC and *Kappa* for $M_1(A)$. Increasing asymmetry but constant entropy.

https://doi.org/10.1371/journal.pone.0222916.g006

We start by introducing a measure of the asymmetry of a matrix $M = (M_{ij})_{i,j=1}^{N}$, say $Asy(M)$, by means of the Frobenius norm of the difference between the matrix and its transpose. That is to say, we define

$$Asy(M) = ||M - M^T|| = \sqrt{\sum_{i,j=1}^{N} (M_{ij} - M_{ji})^2}.$$

Example (a) Let us consider matrix $M_1(A) = \begin{pmatrix} 1 & 2A & A \\ A & 1 & 2A \\ A & A & 1 \end{pmatrix}$, with $A \geq 1$. Obviously, $M_1(A)$ is not symmetric, with $Asy(M_1(A)) = 2A$, which increases with A, achieving the minimum = 2 when $A = 1$. We can make a graph showing the evolution of *Kappa* and MCC when increasing A, as shows Fig 6, where it can be observed that the behaviour of *Kappa* is very similar to that of MCC. Then, asymmetry has not been enough to generate a different behaviour of them. What, then?

Think about the entropy generated by the values of the matrix that are outside the main diagonal. In general, given a set of non-negative numbers, say $\{n_1, \ldots, n_r\}$, the Shannon's entropy generated by the set can be defined by $Ent = \sum_{i=1}^{r} -p_i \log(p_i)$, with $p_i = \frac{n_i}{n}$ if $n = \sum_{i=1}^{r} n_i$, where log usually denotes logarithm in base 2. With this definition, $Ent(M_1(A)) = Ent(\{2A, A, A, 2A, A, A\}) = 2.5$, which is independent of A, so for the family of matrices $M_1(A)$, entropy can not play any role since it remains constant when A varies. The same happens with matrix C_0, for which asymmetry increases as $b \to +\infty$ but entropy remains constant. In other words: increasing asymmetry but constant entropy does not produce the phenomenon of inappropriate behaviour of *Kappa* in which we are interested.

Example (b) Consider now matrix $M_2(A) = \begin{pmatrix} 1 & A & 1 \\ 1 & 1 & A^2 \\ 1 & 1 & 1 \end{pmatrix}$ with $A > 1$. Then

$Asy(M_2(A)) = \sqrt{2}\,(A-1)\sqrt{1 + (A+1)^2}$, which increases with A, and

$Ent(M_2(A)) = Ent(\{A, 1, 1, A^2, 1, 1\}) = \log(A\,(A+1) + 4) - \frac{A\,(2\,A+1)}{A\,(A+1)+4} \log(A)$

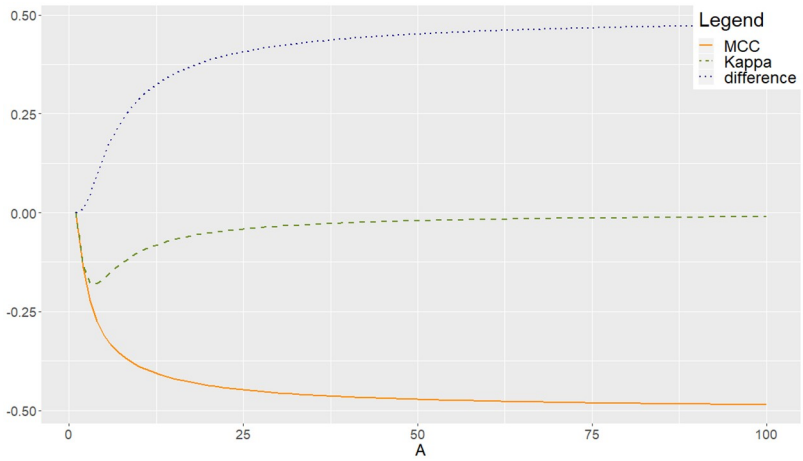

Fig 7. Experimental disagreement between MCC and *Kappa* for $M_2(A)$. Decreasing to zero entropy, which implies increasing asymmetry.

https://doi.org/10.1371/journal.pone.0222916.g007

decreases, converging to 0 as $A \to +\infty$. The corresponding plots of *Kappa*, MCC and the difference, with respect to A are shown in Fig 7.

$MCC(M_2(A))$ is a decreasing function of A but $\mathcal{K}(M_2(A))$ is increasing for $A \geq 4$. Then, we can observe a contradictory behaviour of the two measures. Let us see this with numerical examples in Table 1: as A increases (and then, asymmetry increases while entropy decreases to zero), MCC decreases but *Kappa* increases.

Remark 4 *Note that for matrix* $M_2(A)$, MCC *and Kappa diverge as A increases, as it happens with the family of matrices* Z_A *and with the confusion matrix* $C_1 = \begin{pmatrix} a & b \\ 1 & d \end{pmatrix}$ *considered in Proposition 3 (binary case with* $c = 1$ *in which the behaviour of Kappa appears as contrary to common sense when b increases). In the three scenarios, entropy decreases to zero and the asymmetry of the confusion matrix grows to* $+\infty$. *Indeed, for matrices* Z_A *(as* $A \to +\infty$*) and* C_1 *(as* $b \to +\infty$*) we have that*

$$\begin{aligned} Asy(Z_A) &= \sqrt{2}\,(A-1) \nearrow +\infty\,, \\ Ent(Z_A) &= \log(N^2 - 1 + A) - \frac{A}{N^2 - 1 + A}\log(A) \searrow 0\,, \\ Asy(C_1) &= \sqrt{2}\,|b-1| \nearrow +\infty\,, \\ Ent(C_1) &= Ent(\{1, b\}) = \frac{-b}{b+1}\log(b) + \log(b+1) \searrow 0\,. \end{aligned}$$

In general, entropy of the elements outside the main diagonal and asymmetry are related in the sense given by the following lemma.

Table 1. Comparing MCC, *Kappa*, *Asy* and *Ent* for $M_2(A)$. $A = 10, 25, 50, 75, 100$.

$M_2(A)$	$A = 10$	$A = 25$	$A = 50$	$A = 75$	$A = 100$
MCC	-0.3879	-0.4478	-0.4722	-0.4810	-0.4856
Kappa	-0.1002	-0.0410	-0.0203	-0.0135	-0.0101
Asy	140.5845	883.1217	3534.7990	7954.2260	14141.4100
Ent	0.7135	0.2998	0.1590	0.1108	0.0859

https://doi.org/10.1371/journal.pone.0222916.t001

Lemma 9 *Let* $C(A) = (C_{ij}(A))_{i,j=1,\ldots,N}$ *be a matrix of non-negative integers depending on a parameter* $A \in \mathbb{N}$, *and such that* $Ent(C(A)) > 0$ *for any A. Therefore, if the entropy of* $C(A)$ *decreases to zero, asymmetry must grow to infinity, that is*,

$$\lim_{A\to+\infty} Ent(C(A)) = 0 \quad \Rightarrow \quad \lim_{A\to+\infty} Asy(C(A)) = +\infty\,.$$

Proof: By definition of Shannon's entropy, if $Ent(C(A))$ converges to zero, then in the limit there is no uncertainty outside the main diagonal, that is, there must exist a pair (i, j), with $i \neq j$, such that

$$\lim_{A\to+\infty} C_{ij}(A) = +\infty \quad \text{and} \quad \forall (r,s) \neq (i,j),\ \lim_{A\to+\infty} \frac{C_{rs}(A)}{C_{ij}(A)} = 0\,.$$

Then, with $(r, s) = (j, i)$, we can write

$$\lim_{A\to+\infty} (C_{ij}(A) - C_{ji}(A))^2 = \lim_{A\to+\infty} (1 - \frac{C_{ji}(A)}{C_{ij}(A)})^2\, C_{ij}^2(A) = +\infty$$

since $\lim_{A\to+\infty}(1 - \frac{C_{ji}(A)}{C_{ij}(A)})^2 = (1-0)^2 = 1$ and $\lim_{A\to+\infty} C_{ij}^2(A) = +\infty$.

Finally, from the fact that $Asy(C(A)) \geq |C_{ij}(A) - C_{ji}(A)| \to +\infty$ we finish the proof.

Lemma 9 confirms that what we have observed in different examples (confusion matrices C_1 as function of b, Z_A and $M_2(A)$), in which entropy tended to zero and asymmetry grew towards infinity, is not a coincidence but the rule.

It is still necessary to ask whether the role of asymmetry in observing the phenomenon of the discrepancy between the behaviours of *Kappa* and MCC is canceled out by entropy. That is, if the phenomenon still can be observed if the asymmetry remains constant while the entropy does not decrease to zero. The negative answer is given by the following example, in which asymmetry is constant and entropy decreases to a positive limit but the phenomenon of discrepancy between MCC and *Kappa* is no longer observed.

Example (c) Be matrix $M_3(A) = \begin{pmatrix} 1 & B & B \\ B+100 & 1 & B \\ B+100 & B+100 & 1 \end{pmatrix}$ with $B = 1000 - A$, $A = 0,\ldots,$ 999. The corresponding plot of MCC, *Kappa* and the difference in absolute value is shown in Fig 8. In this setting, as with Example (a), there is an agreement in the behaviour of MCC and *Kappa*. However, in this case there is no decrease of entropy to zero as in Example (b). Indeed, $Ent(M_3(A)) = \log(6\,B + 300) - \frac{1}{6\,B+300}(3\,B\log(B) + 3\,(B+100)\log(B+100))$ with $B = 1000 - A$, is a monotonically decreasing function of A that converges to $\log(300) - \log(100) > 0$ as $A \to 1000$, while $Asy(M_3(A)) = 100\sqrt{6}$ remains constant.

Previous examples, in which the diagonal stays constant, show that it is not enough that the asymmetry grows to infinity, or that the entropy is constant or simply decreasing, for the phenomenon of discrepancy between *Kappa* and MCC to occur, but heuristically it seems that entropy must decrease to zero, which implies that at the same time asymmetry grows to infinity by Lemma 9. At least it is what experimentation has shown in the cases already commented. To finish, two more examples in the same vein, the first corresponding to the situation of discrepancy, and the latter to the similarity, in the behaviours of MCC and Kappa.

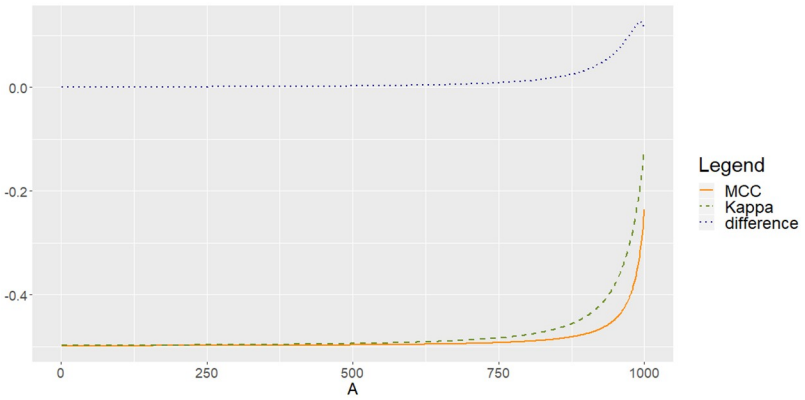

Fig 8. Experimental agreement between MCC and *Kappa* for $M_3(A)$. Decreasing entropy to a positive limit and constant asymmetry.

https://doi.org/10.1371/journal.pone.0222916.g008

Example (d) Let be the confusion matrix $M_4(A) = \begin{pmatrix} 1 & A & 1 \\ A^2 & 1 & B \\ 1 & B^2 & 1 \end{pmatrix}$, with $B = 100 - A$ and $A = 50,\ldots, 100$. In this case, as function of $A \in [50, 100]$,

$$Asy(M_4(A)) = \sqrt{2}\sqrt{A^2(A-1)^2 + (100-A)^2(99-A)^2}$$

monotonically increases with A, and

$$Ent(M_4(A)) = \log(g(A)) - \frac{A(2A+1)\log(A) + (100-A)(201-2A)\log(100-A)}{g(A)},$$

with $g(A) = A(A+1) + (100-A)(101-A) + 2$, monotonically decreases (to zero if we increase the parameter 100). We can observe in Fig 9 that in this case the appearance of the described phenomenon of behaviour against the common sense of *Kappa* is confirmed: for $A > 50$, MCC decreases and *Kappa* increases as A increases. By symmetry, for $A < 50$ we observe just the same when A decreases.

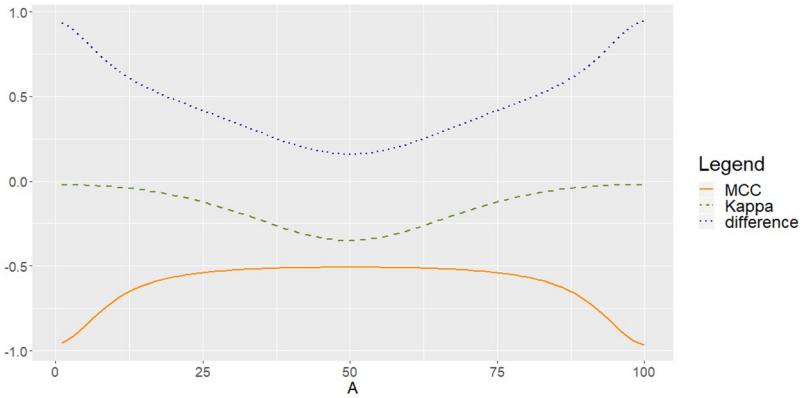

Fig 9. Experimental disagreement between MCC and *Kappa* for $M_4(A)$. Entropy decreases to zero, which implies that asymmetry increases, for A increasing from 50 to 100, and from A decreasing from 50 to 0, by symmetry.

https://doi.org/10.1371/journal.pone.0222916.g009

Table 2. Comparing MCC, *Kappa*, *Asy* and *Ent* for $M_4(A)$. A = 50, 60, 70, 80, 90, 100.

$M_4(A)$	$A = 50$	$A = 60$	$A = 70$	$A = 80$	$A = 90$	$A = 100$
MCC	-0.5081	-0.5114	-0.5249	-0.5653	-0.7032	-0.9659
Kappa	-0.3500	-0.2900	-0.1735	-0.0817	-0.0341	-0.0200
Asy	4900.0000	5470.868	6940.576	8953.971	11328.5700	14000.7100
Ent	1.1442	1.0319	0.7554	0.4418	0.1970	0.0830

https://doi.org/10.1371/journal.pone.0222916.t002

Table 2 illustrates this example numerically through a particular case in which we compare different values of A. We observe that when entropy decreases and asymmetry increases ($A > 50$) MCC decreases and *Kappa* increases, while a completely symmetrical behaviour is observed for $A < 50$, according to Fig 9.

Example (e) Let be the confusion matrix $M_5(A) = \begin{pmatrix} 1 & 2A & A \\ A & 1 & A+100 \\ A & A & 1 \end{pmatrix}$. As function of $A \geq 1$, $Asy(M_5(A)) = \sqrt{2}\sqrt{A^2 + 100^2} \nearrow +\infty$ and is increasing, while

$$Ent(M_5(A)) = \log(7A + 100) - \frac{8A\log(A) + (A + 100)\log(A + 100)}{7A + 100}$$

decreases to $\log(7) - 2/7 > 0$ when $A \to +\infty$. In this case, MCC and *Kappa* agree in behaviour as A increases.

Conclusion

Accuracy is one of the most intuitive and widely used performance metrics for classification although it is not appropriate when considering unbalanced cases. MCC and *Kappa* seem to correct this bias: the former was initially designed to deal with very unbalanced data, while the latter, which was not created to be a classification performance metric but that, however, is widely used for this, takes into account the probability of getting the classification by pure chance. These two measures have a similar behaviour in some situations. In fact, we show that they coincide precisely when the confusion matrix is perfectly symmetric. In other situations, however, their behaviour can diverge to the point that *Kappa* should be avoided as a measure of behaviour to compare classifiers in favor of more robust measures as MCC.

In the present work, similarities and differences among MCC and *Kappa* have been discussed and illustrated with synthetic confusion matrices, both in the binary and in the multiclass setting. Our mathematical analysis and heuristic study show that in situations in which the diagonal of the confusion matrix stays constant and at the same time there is a decrease to zero of the entropy of the elements outside the diagonal, which implies an increase in the asymmetry of the confusion matrix, the phenomenon of qualitative differentiation in the behaviour of *Kappa* and MCC appears clearly. Notwithstanding, neither increasing nor constant asymmetry when entropy is not decreasing to zero, does not seem to be enough to produce this phenomenon. As far as we know, this kind of conclusions have not been reached before, so they represent a novelty in the study of *Kappa*.

From a clinical perspective, the fact that *Kappa* is a relative measure of agreement is problematic since it is hard to set a threshold for a good agreement. This does not seem to be a problem when it is used as a performance metric, because *Kappa* values are compared for each classifier given a unique ground-truth, being the relative difference and not the value itself, which determines the best classifier. Notwithstanding, we have shown that if marginal

Table 3. Summary of the obtained results: Examples and agreement/disagreement between the behaviour of MCC and *Kappa* in terms of the asymmetry of the confusion matrix and of the entropy associated to the elements outside the main diagonal. Disagreement scenario corresponds to entropy decreasing to zero, which implies by Lemma 9 that asymmetry must grow to infinity.

	Asymmetry $\nearrow +\infty$	Asymmetry = constant
Entropy = constant	Agreement C_0, $M_1(A)$	
Entropy $\searrow 0$	Disagreement C_1, Z_A, $M_2(A)$, $M_4(A)$	
Entropy $\searrow > 0$	Agreement $M_5(A)$	Agreement $M_3(A)$

https://doi.org/10.1371/journal.pone.0222916.t003

probabilities are really small, the distribution of the misclassification also affects the value of *Kappa*, to the extent that worse classification results can obtain, however, higher values of the statistic. This is especially dramatic when the entropy of the elements outside the main diagonal of the confusion matrix decreases to zero.

A summary of the examples that have been considered in this work according to the agreement/disagreement between the behaviour of MCC and *Kappa*, can be found in the Table 3.

The standard problems associated with *Kappa* are mainly related to unbalanced datasets (see for instance [36] and [17]). We show that an unbalanced situation can make *Kappa* not comparable between different situations, but to achieve counter-intuitive results, it is also necessary that the entropy of the elements outside the main diagonal to decrease to zero.

Nowadays, in the field of machine learning such situations, in which the number of observations of one of the classes far exceed the quantity of the others, or when the marginal distributions are small, are very common. Machine learning algorithms automatically scrutinize huge amount of data, classifying it into hundreds of categories or look for an unlikely relevant event. In that framework, the finding of a dependable performance measure to be robust and reliable becomes of the utmost importance. Hence, we believe that it has been sufficiently justified that, unfortunately, Cohen's *Kappa* can no longer play this role, especially considering the existence of solid alternatives.

Acknowledgments

The authors wish to thank the anonymous referees for careful reading and helpful comments that resulted in an overall improvement of the paper.

Author Contributions

Writing – original draft: Rosario Delgado, Xavier-Andoni Tibau.

Writing – review & editing: Rosario Delgado, Xavier-Andoni Tibau.

References

1. Ferri C., Hernández-Orallo J., Modroiu R.: An experimental comparison of performance measures for classification. Pattern Recognition Letters 30(1), 27–38 (2009) https://doi.org/10.1016/j.patrec.2008.08.010
2. Jurman G., Riccadonna S., Furlanello C.: A comparison of mcc and cen error measures in multi-class prediction. PloS one 7(8), e41882 (2012) https://doi.org/10.1371/journal.pone.0041882
3. Sokolova M., Lapalme G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management 45(4), 427–437 (2009) https://doi.org/10.1016/j.ipm.2009.03.002
4. Matthews B.W.: Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure 405(2), 442–451 (1975) https://doi.org/10.1016/0005-2795(75)90109-9

5. Gorodkin J.: Comparing two k-category assignments by a k-category correlation coefficient. Computational biology and chemistry 28(5-6), 367–374 (2004) https://doi.org/10.1016/j.compbiolchem.2004.09.006 PMID: 15556477
6. Stokić D., Hanel R., Thurner S.: A fast and efficient gene-network reconstruction method from multiple over-expression experiments. BMC bioinformatics 10(1), 253 (2009) https://doi.org/10.1186/1471-2105-10-253 PMID: 19686586
7. Supper, J., Spieth, C., Zell, A.: Reconstructing linear gene regulatory networks. In: European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, pp. 270–279. Springer (2007)
8. Blair E., Stanley F.: Interobserver agreement in the classification of cerebral palsy. Developmental Medicine & Child Neurology 27(5), 615–622 (1985) https://doi.org/10.1111/j.1469-8749.1985.tb14133.x
9. Cameron M.L., Briggs K.K., Steadman J.R.: Reproducibility and reliability of the outerbridge classification for grading chondral lesions of the knee arthroscopically. The American journal of sports medicine 31(1), 83–86 (2003) https://doi.org/10.1177/03635465030310012601 PMID: 12531763
10. Monserud R.A., Leemans R.: Comparing global vegetation maps with the Kappa statistic. Ecological modelling 62(4), 275–293 (1992) https://doi.org/10.1016/0304-3800(92)90003-W
11. Allouche O., Tsoar A., & Kadmon R.: Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of applied ecology 43(6), 1223–1232 (2006) https://doi.org/10.1111/j.1365-2664.2006.01214.x
12. Tian Y., Zhang H., Pang Y., Lin J.: Classification for single-trial N170 during responding to facial picture with emotion. Front. Comput. Neurosci. 12:68. https://doi.org/10.3389/fncom.2018.00068 PMID: 30271337
13. Donker D., Hasman A., Van Geijn H.: Interpretation of low Kappa values. International journal of biomedical computing 33(1), 55–64 (1993) PMID: 8349359
14. Forbes A.D.: Classification-algorithm evaluation: Five performance measures based onconfusion matrices. Journal of Clinical Monitoring 11(3), 189–206 (1995) https://doi.org/10.1007/BF01617722 PMID: 7623060
15. Brennan R.L., Prediger D.J.: Coefficient Kappa: Some uses, misuses, and alternatives. Educational and psychological measurement 41(3), 687–699 (1981) https://doi.org/10.1177/001316448104100307
16. Maclure M., Willett W.C.: Misinterpretation and misuse of the Kappa statistic. American journal of epidemiology 126(2), 161–169 (1987) https://doi.org/10.1093/aje/126.2.161 PMID: 3300279
17. Uebersax J.S.: Diversity of decision-making models and the measurement of interrater agreement. Psychological bulletin 101(1), 140–146 (1987) https://doi.org/10.1037/0033-2909.101.1.140
18. Feinstein A.R., Cicchetti D.V.: High agreement but low Kappa: I. the problems of two paradoxes. Journal of clinical epidemiology 43(6), 543–549 (1990) https://doi.org/10.1016/0895-4356(90)90158-l PMID: 2348207
19. Cicchetti D.V., Feinstein A.R.: High agreement but low Kappa: Ii. resolving the paradoxes. Journal of clinical epidemiology 43(6), 551–558 (1990) https://doi.org/10.1016/0895-4356(90)90159-m PMID: 2189948
20. Krippendorff K.: Reliability in content analysis: Some common misconceptions and recommendations. Human communication research 30(3), 411–433 (2004) https://doi.org/10.1111/j.1468-2958.2004.tb00738.x
21. Warrens M.J.: A formal proof of a paradox associated with Cohen's Kappa. Journal of Classification 27 (3), 322–332 (2010) https://doi.org/10.1007/s00357-010-9060-x
22. Byrt T., Bishop J., & Carlin J. B.: Bias, prevalence and kappa. Journal of clinical epidemiology 46(5), 423–429 (1993) https://doi.org/10.1016/0895-4356(93)90018-v PMID: 8501467
23. de Vet H.C., Mokkink L.B., Terwee C.B., Hoekstra O.S., Knol D.L.: Clinicians are right not to like Cohen's Kappa. BMJ 346, f2125 (2013) https://doi.org/10.1136/bmj.f2125 PMID: 23585065
24. Dice L. R.: Measures of the amount of ecologic association between species. Ecology 26(3), 297–302 (1945) https://doi.org/10.2307/1932409
25. Albatineh A. N., Niewiadomska-Bugaj M., & Mihalko D.: On similarity indices and correction for chance agreement. Journal of Classification 23(2), 301–313 (2006) https://doi.org/10.1007/s00357-006-0017-z
26. Warrens M. J.: On similarity coefficients for 2 × 2 tables and correction for chance. Psychometrika 73 (3), 487 (2008) https://doi.org/10.1007/s11336-008-9059-y PMID: 20037641
27. Cohen J.: A coefficient of agreement for nominal scales. Educational and psychological measurement 20(1), 37–46 (1960) https://doi.org/10.1177/001316446002000104

28. Scott W.A.: Reliability of content analysis: The case of nominal scale coding. Public opinion quarterly pp. 321–325 (1955) https://doi.org/10.1086/266577
29. Mak T. K.: Analysing intraclass correlation for dichotomous variables. Journal of the Royal Statistical Society: Series C (Applied Statistics) 37(3), 344–352 (1988)
30. Goodman L. A., & Kruskal W. H.: Measures of association for cross classifications III: Approximate sampling theory. Journal of the American Statistical Association, 58(302), 310–364 (1963) https://doi.org/10.1080/01621459.1963.10500850
31. Brennan R. L., & Light R. J.: Measuring agreement when two observers classify people into categories not defined in advance. British Journal of Mathematical and Statistical Psychology 27(2), 154–163 (1974) https://doi.org/10.1111/j.2044-8317.1974.tb00535.x
32. Bexkens R., Claessen F. M., Kodde I. F., Oh L. S., Eygendaal D., & van den Bekerom M. P.: The kappa paradox. Shoulder & Elbow, 10(4), 308–308 (2018) https://doi.org/10.1177/1758573218791813
33. Viera A. J., & Garrett J. M.: Understanding interobserver agreement: the kappa statistic. Fam med 37 (5), 360–363 (2005) PMID: 15883903
34. Sim J., & Wright C. C.: The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical therapy 85(3), 257–268 (2005) PMID: 15733050
35. Warrens M.J.: On association coefficients, correction for chance, and correction for maximum value. Journal of Modern Mathematics Frontier 2(4), 111–119 (2013)
36. Andrés A.M., Marzo P.F.: Delta: A new measure of agreement between two raters. British journal of mathematical and statistical psychology 57(1), 1–19 (2004) https://doi.org/10.1348/000711004849268 PMID: 15171798
37. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research 12(Oct), 2825–2830 (2011)
38. Kuhn M., et al.: Caret package. Journal of statistical software 28(5), 1–26 (2008)
39. Huang C., Davis L., Townshend J.: An assessment of support vector machines for land cover classification. International Journal of remote sensing 23(4), 725–749 (2002) https://doi.org/10.1080/01431160110040323
40. Duro D.C., Franklin S.E., Dubé M.G.: A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using spot-5 HRG imagery. Remote Sensing of Environment 118, 259–272 (2012) https://doi.org/10.1016/j.rse.2011.11.020
41. Passos A.N., Kohara V.S., Freitas R.S.d., Vicentini A.P.: Immunological assays employed for the elucidation of an histoplasmosis outbreak in São Paulo, SP. Brazilian Journal of Microbiology 45(4), 1357–1361 (2014) https://doi.org/10.1590/s1517-83822014000400028 PMID: 25763041
42. Claessen F. M., van den Ende K. I., Doornberg J. N., Guitton T. G., Eygendaal D., van den Bekerom M. P., . . . & Wagener M.: Osteochondritis dissecans of the humeral capitellum: reliability of four classification systems using radiographs and computed tomography. Journal of shoulder and elbow surgery 24 (10), 1613–1618 (2015) https://doi.org/10.1016/j.jse.2015.03.029 PMID: 25953486
43. Powers, D.M.W.: The problem with Kappa. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 345–355. Association for Computational Linguistics (2012)
44. Jeni, L.A., Cohn, J.F., De La Torre, F.: Facing imbalanced data–recommendations for the use of performance metrics. In: Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, pp. 245–251. IEEE (2013)
45. Zhao X., Liu J.S., Deng K.: Assumptions behind intercoder reliability indices. In Salmon Charles T. (ed.) Communication Yearbook 36, 419–480. New York: Routledge (2013)
46. Witten I.H., Frank E., Hall M.A., Pal C.J.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2016)
47. Krippendorff K.: Association, agreement, and equity. Quality and Quantity 21(2), 109–123 (1987) https://doi.org/10.1007/BF00167603
48. Krippendorff K.: Content analysis: An introduction to its methodology (1980)