A Comparison of MCC and CEN Error Measures in Multi-Class Prediction

Giuseppe Jurman*, Samantha Riccadonna, Cesare Furlanello

Fondazione Bruno Kessler, Trento, Italy

Abstract

We show that the Confusion Entropy, a measure of performance in multiclass problems has a strong (monotone) relation with the multiclass generalization of a classical metric, the Matthews Correlation Coefficient. Analytical results are provided for the limit cases of general no-information (n-face dice rolling) of the binary classification. Computational evidence supports the claim in the general case.

Citation: Jurman G, Riccadonna S, Furlanello C (2012) A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. PLoS ONE 7(8): e41882. doi:10.1371/journal.pone.0041882

Editor: Giuseppe Biondi-Zoccai, Sapienza University of Rome, Italy

Received January 31, 2012; **Accepted** June 27, 2012; **Published** August 8, 2012

Funding: The authors acknowledge funding by the European Union FP7 Project HiperDART and by the PAT funded Project ENVIROCHANGE. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: jurman@fbk.eu

Introduction

Comparing classifiers' performance is one of the most critical tasks in machine learning. Comparison can be carried out either by means of statistical tests [1,2] or by adopting a performance measure as an indicator to derive similarities and differences, in particular as a function of the number of classes, class imbalance, and behaviour on randomized labels [3].

The definition of performance measures in the context of multiclass classification is still an open research topic as recently reviewed [4,5]. One challenging aspect is the extension of such measures from binary to multiclass tasks [6]. Graphical comparison approaches have been introduced [7], but a generic analytic treatment of the problem is still unavailable.

One relevant case study regards the attempt of extending the Area Under the Curve (AUC) measure, which is one of the most widely used measures for binary classifiers but it has no automatic extension to the multiclass case. The AUC is associated to the Receiver Operating Characteristic (ROC) curve [8,9] and thus proposed formulations were based on a multiclass ROC approximation [10–13]. A second class of extensions is defined by the Volume Under the Surface (VUS) approach, which is obtained by considering the generalized ROC as a surface whose volume has to be computed by exact integration or polynomial approximation [14–16]. As a baseline, the average of the AUCs on the pairwise binary problems derived from the multi-class problems has also been proposed [17].

Other measures are more naturally extended, such as the accuracy (ACC, *i.e.* the fraction of correctly predicted samples), the Global Performance Index [18,19], and the Matthews Correlation Coefficient (MCC). We will focus our attention to the last function [20], which in the binary case is also known as the ϕ-coefficient, *i.e.*, the square root of the average χ^2 statistic $\sqrt{\chi^2/n}$ on n observed samples for the 2×2 contingency table of the classification problem.

For binary tasks, MCC has attracted the attention of the machine learning community as a method that summarizes into a single value the confusion matrix [21]. Its use as a reference performance measure on unbalanced data sets is now common in other fields such as bioinformatics. Remarkably, MCC was chosen as accuracy index in the US FDA-led initiative MAQC-II for comparing about 13 000 different models, with the aim of reaching a consensus on the best practices for development and validation of predictive models based on microarray gene expression and genotyping data [22]. A generalization of MCC to the multiclass case was defined in [23], also used for comparing network topologies [24,25].

A second family of measures that have a natural definition for multiclass confusion matrices are the functions derived from the concept of (information) Entropy, first introduced in [26]. In the classification framework, measures in the entropy family range from the simpler confusion matrix entropy [27] to more complex functions as the Transmitter Information [28] and the Relative Classifier Information (RCI) [29]. Wei and colleagues recently introduced a novel multiclass measure under the name of Confusion Entropy (CEN) [30,31]. They compared CEN to both RCI and accuracy, obtaining better discriminative power and precision in terms of two statistical indicators called degree of consistency and degree of discriminancy [32].

In our study, we investigate the intriguing similarity existing between CEN and MCC. In particular, we experimentally show that the two measures are strongly correlated, and that their relation is globally monotone and locally almost linear. Moreover, we provide a brief outline of the mathematical links between CEN and MCC with detailed examples in limit cases. Discriminancy and consistency ratios are discussed as comparative factors, together with functions of the number of classes, class imbalance, and behaviour on randomized labels.

Methods

Given a classification problem on S samples $\mathcal{S}=\{s_i : 1\leq i\leq S\}$ and N classes $\{1,\ldots,N\}$, define the two functions $\mathrm{tc},\mathrm{pc} : \mathcal{S}\rightarrow\{1,\ldots,N\}$ indicating for each sample s its true class $\mathrm{tc}(s)$ and its predicted class $\mathrm{pc}(s)$, respectively. The corresponding confusion matrix is the square matrix $C\in\mathcal{M}(N\times N,\mathbb{N})$ whose ij-th entry C_{ij} is the number of elements of true class i that have been assigned to class j by the classifier:

$$C_{ij}=|\{s\in\mathcal{S} : \mathrm{tc}(s)=i \textit{ and } \mathrm{pc}(s)=j\}|.$$

The most natural performance measure is the accuracy, defined as the ratio of the correctly classified samples over all the samples:

$$ACC=\frac{\sum_{k=1}^{N} C_{kk}}{S}=\frac{\sum_{k=1}^{N} C_{kk}}{\sum_{i,j=1}^{N} C_{ij}}.$$

Confusion Entropy (CEN)

In information theory, the entropy H associated to a random variable X is the expected value of the self-information $I(X)=-\log p(X)$:

$$H(X)=\mathbb{E}(I(X))=\sum_{x\in X} p(x)I(x)=-\sum_{x\in X} p(x)\log_b(p(x))=\sum_{x\in X} h_b(x),$$

where $p(x)$ is the probability mass function of X, with $h_b(x)=-p(x)\log_b(p(x))=0$ for $p(x)=0$, motivated by the limit $\lim_{x\rightarrow 0} x\log(x)=0$.

The Confusion Entropy measure CEN for a confusion matrix C is defined in [30] as:

$$\begin{aligned}\mathrm{CEN}&=\sum_{j=1}^{N} P_j \mathrm{CEN}_j\\ &=\sum_{j=1}^{N} P_j \sum_{\substack{k=1\\ k\neq j}}^{N} h_{2(N-1)}\left(P^j_{jk}\right)+h_{2(N-1)}\left(P^j_{kj}\right),\end{aligned}\tag{1}$$

where P_j, P^j_{ij}, P^i_{ij} are defined as follows:

P_j is the confusion probability of class j: $P_j=\frac{\sum_{k=1}^{N}\left(C_{jk}\ +C_{kj}\right)}{2\sum_{k,l=1}^{N} C_{kl}}$

P^j_{ij} is the probability of classifying the samples of class i to class j for $i\neq j$ subject to class j:

$$P^j_{ij}=\frac{C_{ij}}{\sum_{k=1}^{N}\left(C_{jk}\ +C_{kj}\right)}$$

P^i_{ij} is the probability of classifying the samples of class i to class j subject to class i:

$$P^i_{ij}=\frac{C_{ij}}{\sum_{k=1}^{N}\left(C_{ik}\ +C_{ki}\right)}\quad \text{for } i\neq j \text{ and } P^i_{ii}=0.$$

For $N>2$, this measure ranges between 0 (perfect classification) and 1 for the complete misclassification case

$$C_{ij}=(1-\delta_{ij})F=\begin{cases}0 & \text{for } i=j\\ F & \text{for } i\neq j \quad\text{and}\quad F\in\mathbb{N}\end{cases},$$

while in the binary case CEN can be greater than 1, as shown below.

Matthews Correlation Coefficient (MCC)

The definition of the MCC in the multiclass case was originally reported in [23]. We recall here the main concepts. Let $X,Y\in\mathcal{M}(S\times N,\mathbb{F}_2)$ be two matrices where $X_{sn}=1$ if the sample s is predicted to be of class n ($\mathrm{pc}(s)=n$) and $X_{sn}=0$ otherwise, and $Y_{sn}=1$ if sample s belongs to class n ($\mathrm{tc}(s)=n$) and 0 otherwise. Using Kronecker's delta function, the definition becomes:

$$X=\left(\delta_{pc(s),n}\right)_{sn}\qquad Y=\left(\delta_{\mathrm{tc}(s),n}\right)_{sn}.$$

Note that $S=\sum_{k,l=1}^{N} C_{kl}$, where $C_{kk}=|\{s\in\mathcal{S} : X_{sk}=Y_{sk}=1\}|=\sum_{s=1}^{S} X_{sk}Y_{sk}$, and, for $k\neq l$, $C_{kl}=|\{s\in\mathcal{S} : X_{sk}=1 \text{ and } Y_{sl}=1\}|$.

The covariance function between X and Y can be written as follows:

$$\begin{aligned}\mathrm{cov}(X,Y)\quad&=\sum_{k=1}^{N} w_k\,\mathrm{cov}(X_k,Y_k)\\ &=\frac{1}{N}\sum_{s=1}^{S}\sum_{k=1}^{N}(X_{sk}-\overline{X}_k)(Y_{sk}-\overline{Y}_k)\end{aligned}$$

where $w_k=\frac{1}{N}$ and $\overline{X}_k$ and $\overline{Y}_k$ are the means of the $k-th$ columns defined respectively as $\overline{X}_k=\frac{1}{S}\sum_{s=1}^{S} X_{sk}=\frac{1}{S}\sum_{l=1}^{N} C_{kl}$ and $\overline{Y}_k=\frac{1}{S}\sum_{s=1}^{S} Y_{sk}=\frac{1}{S}\sum_{l=1}^{N} C_{lk}$.

Finally the Matthews Correlation Coefficient MCC can be written as:

$$\begin{aligned}\mathrm{MCC}\quad&=\frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{cov}(X,X)\cdot\mathrm{cov}(Y,Y)}}\\ &=\frac{\sum_{k,l,m=1}^{N} C_{kk}C_{ml}-C_{lk}C_{km}}{\sqrt{\sum_{k=1}^{N}\left[\left(\sum_{l=1}^{N} C_{lk}\right)\left(\sum_{f,g=1 f\neq k}^{N} C_{gf}\right)\right]}\sqrt{\sum_{k=1}^{N}\left[\left(\sum_{l=1}^{N} C_{kl}\right)\left(\sum_{f,g=1 f\neq k}^{N} C_{fg}\right)\right]}}\end{aligned}\tag{2}$$

MCC lives in the range $[-1,1]$, where 1 is perfect classification. The value -1 is asymptotically reached in the extreme misclassification case of a confusion matrix C with all zeros but

Box 1

$(a)\ S = 60$

$$C = \begin{pmatrix} 15 & 0 & 0 & 0 \\ 0 & 15 & 0 & 0 \\ 0 & 0 & 15 & 0 \\ 0 & 0 & 0 & 15 \end{pmatrix}$$

$\mathrm{MCC} = 1$
$\mathrm{CEN} = 0$

$(b)\ S = 60$

$$C = \begin{pmatrix} 0 & 5 & 5 & 5 \\ 5 & 0 & 5 & 5 \\ 5 & 5 & 0 & 5 \\ 5 & 5 & 5 & 0 \end{pmatrix}$$

$\mathrm{MCC} = -0.333$
$\mathrm{CEN} = 1$

$(c)\ S = 60$

$$C = \begin{pmatrix} 0 & 15 & 0 & 0 \\ 0 & 15 & 0 & 0 \\ 0 & 15 & 0 & 0 \\ 0 & 15 & 0 & 0 \end{pmatrix}$$

$\mathrm{MCC} = 0$
$\mathrm{CEN} = 0.337$

$(d)\ S = 10\ 002$

$$C = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 5000 & 0 \\ 0 & 5000 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$\mathrm{MCC} = -0.999$
$\mathrm{CEN} = 0.387$

Box 2

$(a)\ S = 9, K = 3$

$$C = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

$\mathrm{MCC} = 0$
$\mathrm{CEN} = \frac{2}{3}\log_4 6 = 0.86$

$(b)\ S = 16, K = 4$

$$C = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

$\mathrm{MCC} = 0$
$\mathrm{CEN} = \frac{3}{4}\log_6 8 = 0.87$

$(c)\ S = 25, K = 5$

$$C = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$\mathrm{MCC} = 0$
$\mathrm{CEN} = \frac{4}{5}\log_8 10 = 0.89$

$(d)\ S = 36, K = 6$

$$C = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$\mathrm{MCC} = 0$
$\mathrm{CEN} = \frac{5}{6}\log_{10} 12 = 0.90$

Box 3

$(a)\ S = 25$

$$Z_{10} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 10 & 1 & 1 & 1 \end{pmatrix}$$

$\mathrm{ACC} = 0.160$
$\mathrm{MCC} = -0.088$
$\mathrm{CEN} = 0.713$

$(b)\ S = 115$

$$Z_{100} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 100 & 1 & 1 & 1 \end{pmatrix}$$

$\mathrm{ACC} = 0.035$
$\mathrm{MCC} = -0.154$
$\mathrm{CEN} = 0.207$

$(c)\ S = 1\ 015$

$$Z_{1000} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1000 & 1 & 1 & 1 \end{pmatrix}$$

$\mathrm{ACC} = 0.004$
$\mathrm{MCC} = -0.165$
$\mathrm{CEN} = 0.030$

$(d)\ S = 10\ 015$

$$Z_{10000} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 10000 & 1 & 1 & 1 \end{pmatrix}$$

$\mathrm{ACC} = 4 \cdot 10^{-4}$
$\mathrm{MCC} = -0.167$
$\mathrm{CEN} = 0.004$

Box 4

$S = 12$

(a)

$$C = \begin{pmatrix} 6 & 0 \\ 0 & 6 \end{pmatrix}$$

$\mathrm{MCC} = 1$
$\mathrm{CEN} = 0$

(b)

$$C = \begin{pmatrix} 5 & 1 \\ 1 & 5 \end{pmatrix}$$

$\mathrm{MCC} = 0.67$
$\mathrm{CEN} = 0.60$

(c)

$$C = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$$

$\mathrm{MCC} = 0.33$
$\mathrm{CEN} = 0.86$

(d)

$$C = \begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix}$$

$\mathrm{MCC} = 0$
$\mathrm{CEN} = 1$

(e)

$$C = \begin{pmatrix} 2 & 4 \\ 4 & 2 \end{pmatrix}$$

$\mathrm{MCC} = -0.33$
$\mathrm{CEN} = 1.06$

(f)

$$C = \begin{pmatrix} 1 & 5 \\ 5 & 1 \end{pmatrix}$$

$\mathrm{MCC} = -0.67$
$\mathrm{CEN} = 1.05$

(g)

$$C = \begin{pmatrix} 0 & 6 \\ 6 & 0 \end{pmatrix}$$

$\mathrm{MCC} = -1$
$\mathrm{CEN} = 1$

Figure 1. Examples of CEN and MCC for different confusion matrices.

doi:10.1371/journal.pone.0041882.g001

in two symmetric entries $C_{\bar{i},\bar{j}}$, $C_{\bar{j},\bar{i}}$. MCC is equal to 0 when C is all zeros but for one column (all samples have been classified to be of a class k), or when all entries are equal $C_{ij}=F\in\mathbb{N}$.

Relationships between CEN and MCC

As discussed before, CEN and MCC live in different ranges, whose extreme values are differently reached. In Box 1 of Fig. 1, numerical examples are shown for $K=4$ in different situations: (a) complete classification, (b) complete misclassification, (c) all samples classified as belonging to one class, (d) misclassification case in a very unbalanced situation.

It is worth noting that CEN is more discriminant than MCC in specific situations, although the property is not always welcomed. For instance, in Fig. 1, Box 1(c), $\mathrm{MCC}=0$ while $\mathrm{CEN}\approx 0.337$. Furthermore, as shown in Box 2, $\mathrm{MCC}=0$ for constant matrix $C_{ij}=F\in\mathbb{N}$ for each i,j, regardless of the number of classes N, while it is easy to show that $\mathrm{CEN}=\left(1-\frac{1}{N}\right)\log_{2N-2}2N$, *i.e.*, CEN is a function of N. Note that both measures are invariant for scalar multiplication of the whole confusion matrix, so we always set $F=1$ in Box 2.

For small sample sizes, we can show that CEN has higher discriminant power than MCC, *i.e.*, different confusion matrices can have same MCC and different CEN. This can be quantitatively assessed by using the degree of discriminancy criterion [32]: for two measures f and g on a domain Ψ, let $P=\{(a,b)\in\Psi\times\Psi : f(a)>f(b), g(a)=g(b)\}$ and $Q=\{(a,b)\in\Psi\times\Psi : f(a)=f(b), g(a)>g(b)\}$; then the degree of discriminancy for f over g is $|P|/|Q|$. For instance, as in [30], we consider a 3-class case with 2,4,3 samples respectively: we evaluate all the possible confusion matrices ranging from the perfect classification case

$$\begin{pmatrix}2 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 3\end{pmatrix}$$

to the complete misclassification case. In this case the degree of discriminancy of CEN over MCC is about 6. Similar results hold for all the 12 small sample size cases on three classes listed in Tab. 6 of [30], ranging from 9 to 19 samples.

We proceed now to show an intriguing relationship between MCC and CEN. First consider the confusion matrix B of dimension N where $B_{ji}=F+(T-F)\delta_{ij}$, *i.e.*, all entries have value F but in the diagonal whose values are all T, for T, F two integers. In this case,

$$\begin{aligned}\mathrm{MCC} &= \frac{T^2+(N-2)TF-(N-1)F^2}{[T+(N-1)F]^2},\\ \mathrm{CEN} &= \frac{(N-1)F}{T+(N-1)F}\log_{2N-2}\frac{2[T+(N-1)F]}{F},\end{aligned}$$

and thus

$$\mathrm{CEN}=(1-\mathrm{MCC})\left(1+\log_{2N-2}\frac{T+(N-1)F}{(N-1)F}\right)\left(1-\frac{1}{N}\right).$$

This identity can be relaxed to the following generalization, which slightly underestimates CEN:

$$\begin{aligned}\mathrm{CEN} &\simeq \frac{1}{k}\cdot(1-\mathrm{MCC})\left(1+\log_{2N-2}\frac{\sum_{i,j=1}^{N}C_{ij}}{\sum_{i,j=1\, i\neq j}^{N}C_{ij}}\right)\\ &\left(1-\frac{1}{N}\right)\\ &=\frac{1}{k}\cdot(1-\mathrm{MCC})[1-\log_{2N-2}(1-\mathrm{ACC})]\left(1-\frac{1}{N}\right)\end{aligned} \tag{3}$$

where both sides are zero when $\mathrm{MCC}=\mathrm{ACC}=1$, and $k\approx 1.012\cdot\left(1+\frac{0.18924}{\log(N)}-\frac{0.06694}{\log^2(N)}\right)$. For simplicity's sake, we call "transformed MMC" (tMCC) the right member of Eq. 3.

A numerical simulation shows that the tMCC approximation in Eq. 3 holds in a more general and practical setting (Fig. 2). In the simulation, 200 000 confusion matrices M_i (dimension range: 3 to 30) were generated. For each class j, the number of correctly classified elements (*i.e.*, the j-th diagonal element) was uniformly randomly chosen between 1 and 1000. Then the off-diagonal entries were generated as random integers between 1 and $\lfloor 1000\rho_i \rfloor$, where the parameter ρ_i was extracted from the uniform distribution in the range $[0.01,1]$, corresponding to small-moderate misclassification. For such data, the Pearson correlation between tMCC and $k\cdot$CEN is about 0.994.

In order to compare measures, we consider also the degree of consistency indicator [32]: for two measures f and g on a domain Ψ, let $R=\{(a,b)\in\Psi\times\Psi : f(a)>f(b), g(a)>g(b)\}$ and $V=\{(a,b)\in\Psi\times\Psi : f(a)>f(b), g(a)<g(b)\}$; then the degree of consistency c of f and g is $c(f,g)=|R|/(|R|+|V|)$. On the given data, $c(\mathrm{tMCC},k\cdot\mathrm{CEN})\approx 1-10^{-7}$, while the degree of discriminancy is undefined since no ties occur. In summary, the relation between tMMC and $k\cdot$CEN is close to linear on this data, with an average ratio of 1.000508 (CI: $1.000328-1.000711$, 95% bootstrap Student).

Comparison on the Z_{A} family. The behaviour of the Confusion Entropy is instead rather diverse from MCC and ACC for the family of Z_A matrices, where all entries are equal but for a non-diagonal one. Because of the multiplicative invariance, all entries can be set to one but for the leftmost lower corner: $(Z_A)_{ij}=1+\delta_{(i,j),(N,1)}(A-1)$ for $A\geq 1$ a positive integer. As shown in Fig. 1, Box 3, when A grows bigger, more and more samples are misclassified, *i.e.*, the accuracy $\mathrm{ACC}(Z_A)=N/(N^2+A-1)$ decreases to zero for increasing A.

The MCC measure of this confusion matrix is

$$\mathrm{MCC}(Z_A)=-\frac{A-1}{(N-1)(N^2+2A-2)},$$

which is a function monotonically decreasing for increasing values of A, with limit $-\frac{1}{2(N-1)}$ for $A\rightarrow\infty$.

On the other hand, the Confusion Entropy for the same family of matrices is

$$\begin{aligned}\mathrm{CEN}(Z_A) &= \frac{1}{N^2+A-1}[(N-2)(N-1)\log_{2N-2}(2N)\\ &+(2N+A-3)\log_{2N-2}(2N+A-1)-A\log_{2N-2}(A)],\end{aligned}$$

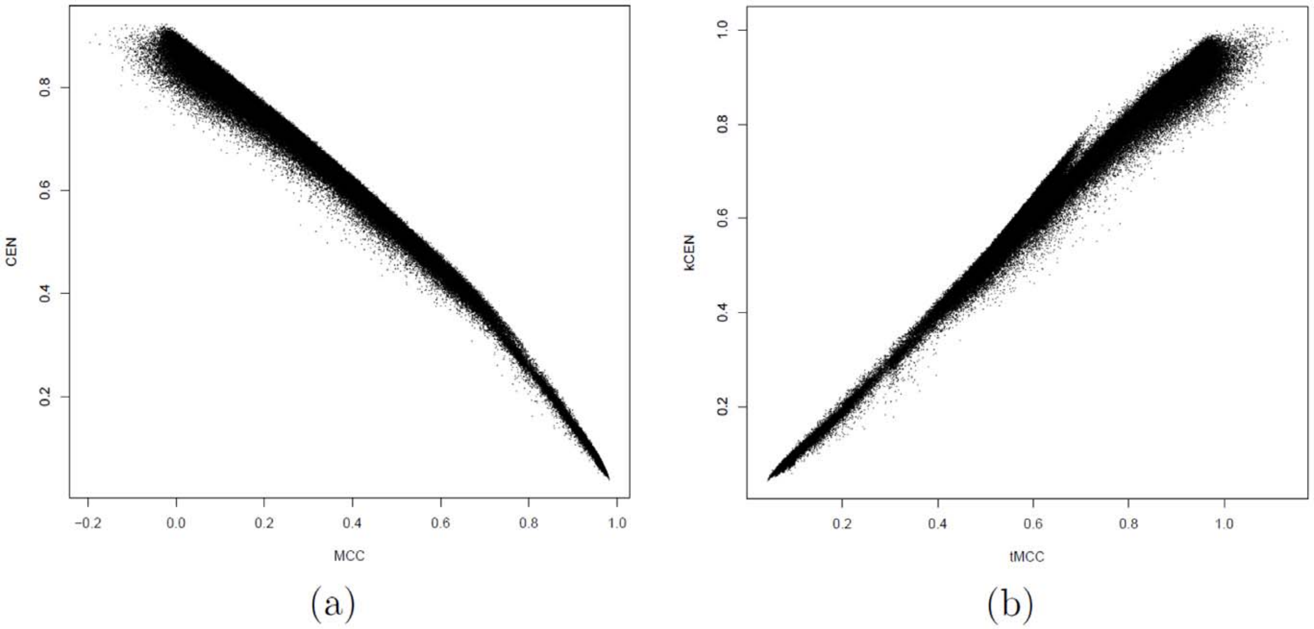


Figure 2. Dotplots of CEN versus MCC. (a) and $k\cdot$CEN versus tMCC (b) for 200 000 random confusion matrices of different dimensions.
doi:10.1371/journal.pone.0041882.g002

which is still a decreasing function of increasing A, but asymptotically moving towards zero, *i.e.*, to the minimal entropy case. In Box 3 of Fig. 1 we present three numerical examples for $A=10,100,1000$.

The dice rolling case. Another pathologic case is found in the case of dice rolling classification on unbalanced classes: because of the multiplicative invariance of the measures, we can assume that the confusion matrix for this case has all entries equal to one but for the last row, whose entries are all A, for $A\geq 1$. In this case, the Confusion Entropy is

$$\begin{aligned}\mathrm{CEN} \quad &= \frac{N-1}{2N(N+A-1)}[(2N+A-3)\log_{2N-2}(2N+A-1)\\ &\quad -2A\log_{2N-2}A+(A+1)\log_{2N-2}(N+NA+A-1)],\end{aligned}$$

a decreasing function for growing A whose limit for $A\to\infty$ is $\frac{N-1}{2N}\log_{2N-2}(N+1)$. As a function of N, this limit is an increasing function asymptotically growing towards $1/2$. It is easy to see that $\mathrm{MCC}\to 0$ for $A\to\infty$ in this case. More in general, while $\mathrm{MCC}=0$ in all those cases where random classification (*i.e.*, no learning) happens, this is lost in the case of CEN, due to its greater discriminant power: there is no unique value associated to the spectrum of random classification problems.

The binary case. In the two-class case (P: positives, N: negatives), the confusion matrix is $\begin{pmatrix}\mathrm{TP} & \mathrm{FN}\\ \mathrm{FP} & \mathrm{TN}\end{pmatrix}$, where T and F stand for true and false respectively. The Matthews Correlation Coefficient has the familiar definition [20,21]:

$$\mathrm{MCC}=\frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}.$$

The Confusion Entropy can be written for the binary case as:

$$\begin{aligned}\mathrm{CEN}=&\frac{(\mathrm{FN}+\mathrm{FP})\log_2((\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})^2-(\mathrm{TP}-\mathrm{TN})^2)}{2(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})}\\ &-\frac{\mathrm{FN}\log_2\mathrm{FN}+\mathrm{FP}\log_2\mathrm{FP}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}.\end{aligned}$$

Note that in the case $\mathrm{TP}=\mathrm{TN}=T\in\mathbb{N}$ and $\mathrm{FP}=\mathrm{FN}=F\in\mathbb{N}$, we have

$$\mathrm{CEN}=\frac{F}{T+F}\log_2\frac{2(T+F)}{F},$$

and thus $\mathrm{CEN}>1$ when the ratio T/F is smaller than 1. In other words, the confusion matrices $\begin{pmatrix}T & F\\ F & T\end{pmatrix}$ with $0<T<F$ have

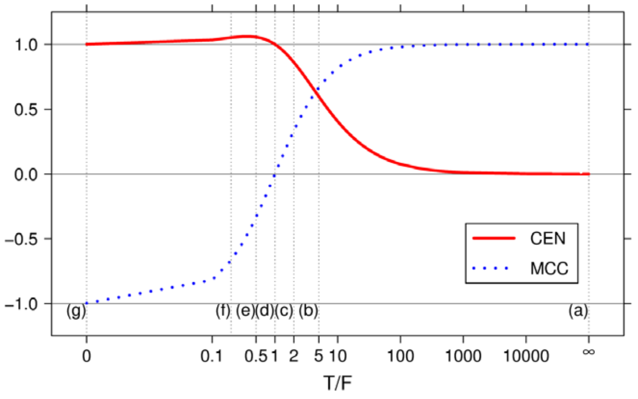
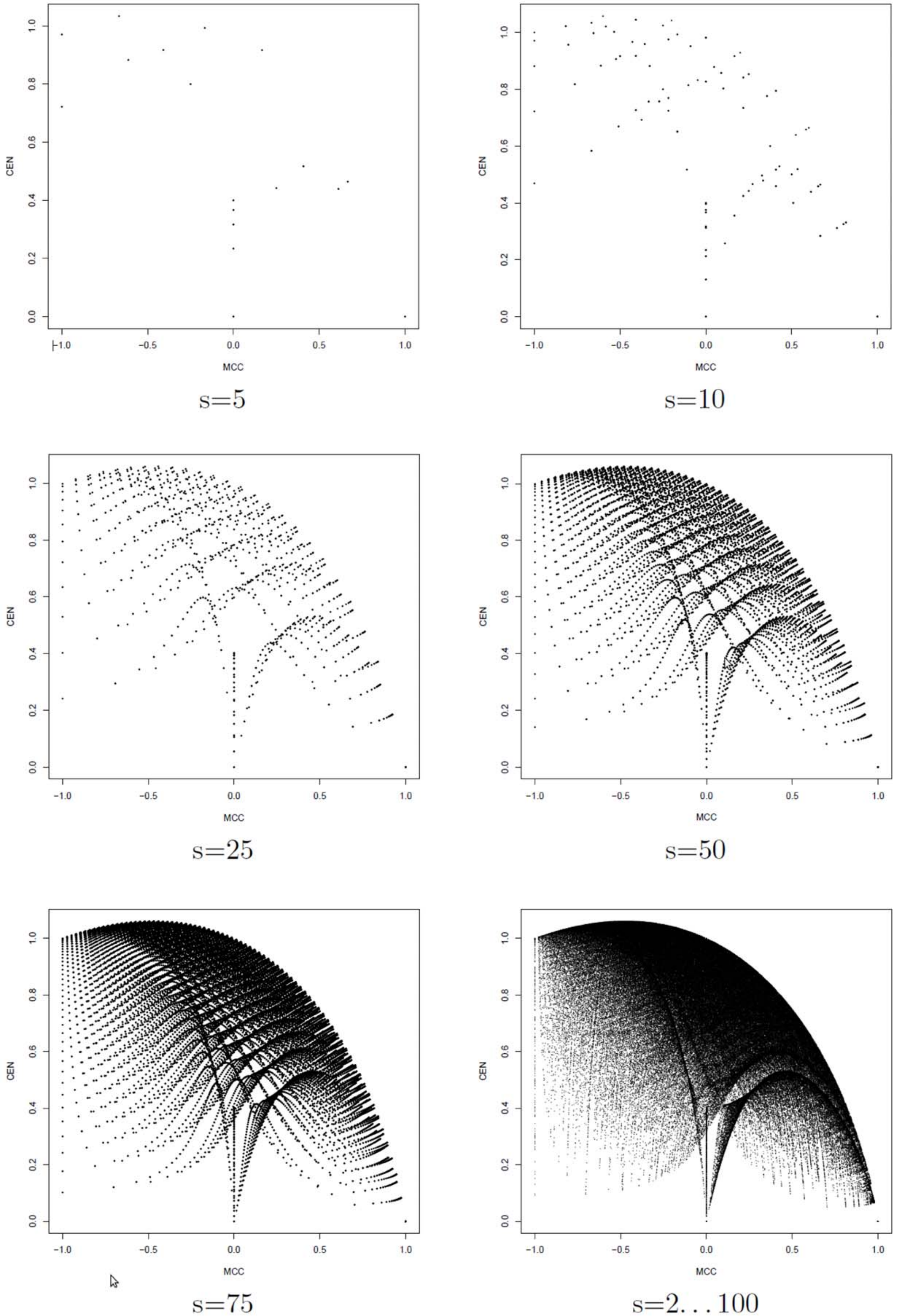


Figure 3. Lines describing CEN and MCC of a confusion matrix $\begin{pmatrix}T & F\\ F & T\end{pmatrix}$ for increasing ratio $\frac{T}{F}$. Gray vertical lines correspond to the examples provided in Fig. 1, Box 4.
doi:10.1371/journal.pone.0041882.g003



s=5

s=10

s=25

s=50

s=75

s=2...100

Figure 4. Scatter plots of CEN versus MCC for all the confusion matrices C_s of binary classification tasks with s = 5,10,50,75 samples and for the cumulative set of all 4 598 125 C_s matrices with $2 \leq s \leq 100$.
doi:10.1371/journal.pone.0041882.g004

$\mathrm{CEN} > 1$; the bound is attained for $T = 0$, the case of total misclassification. This suggests that CEN should not be used as a classifier performance measure in the binary case. A numerical example is provided in Fig. 1, Box 4, while a plot of CEN and MCC curves for different ratios of T/F is shown in Fig. 3.

Indeed, differently from the multi-class case, CEN and MCC are poorly correlated for two classes. We computed MCC and CEN for all the 4 598 125 possible confusion matrices for a binary classification task on s samples ($2 \leq s \leq 100$). Results are displayed in Fig. 4, for $s = 5,10,25,50,75$ and the cumulative plot with all $2 \leq s \leq 100$. In this last case, the (absolute) Pearson correlation between the two metrics is only $\rho \approx 0.63$.

Results and Discussion

We compared the Matthews Correlation Coefficient (MCC) and Confusion Entropy (CEN) as performance measures of a classifier in multiclass problems. We have shown, both analytically and empirically, that they have a consistent behaviour in practical cases. However each of them is better tailored to deal with different situations, and some care should be taken in presence of limit cases.

Both MCC and CEN improve over Accuracy (ACC), by far the simplest and widespread measure in the scientific literature. The point with ACC is that it poorly copes with unbalanced classes and it cannot distinguish among different misclassification distributions.

CEN has been recently proposed to provide an high level of discrimination even between very similar confusion matrices. However, we show that this feature is not always welcomed, as in the case of random dice rolling, for which $\mathrm{MCC} = 0$, but a range of different values is found for CEN. This case is of practical interest because class labels are often randomized as a sanity check in complex classification studies, *e.g.*, in medical diagnosis tasks such as cancer subtyping [33] or image classification problems (*e.g.*, handwritten ZIP code identification or image scene classification examples) [34].

Our analysis also shows that CEN should not be reliably used in the binary case, as its definition attributes high entropy even in regimes of high accuracy and it even gets values larger than one.

In the most general case, MCC is a good compromise among discriminancy, consistency and coherent behaviors with varying number of classes, unbalanced datasets, and randomization. Given the strong linear relation between CEN and a logarithmic function of MCC, they are exchangeable in a majority of practical cases. Furthermore, the behaviour of MCC remains consistent between binary and multiclass settings.

Our analysis does not regard threshold classifiers; whenever a ROC curve can be drawn, generalized versions of the Area Under the Curve algorithm or other similar measures represent a more immediate choice [35]. This given, for confusion matrix analysis, our results indicate that the MCC remains an optimal off-the-shelf tool in practical tasks, while refined measures such as CEN should be reserved for specific topic where high discrimination is crucial.

Author Contributions

Conceived and designed the experiments: GJ SR. Performed the experiments: GJ SR. Analyzed the data: GJ SR. Contributed reagents/materials/analysis tools: GJ SR. Wrote the paper: GJ CF SR.

References

1. Demšar J (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7: 1–30.
2. García S, Herrera F (2008) An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. Journal of Machine Learning Research 9: 2677–2694.
3. Hand D (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning 77: 103–123.
4. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Information Processing and Management 45: 427–437.
5. Ferri C, Hernández-Orallo J, Modroiu R (2009) An experimental comparison of performance measures for classification. Pattern Recognition Letters 30: 27–38.
6. Felkin M (2007) Comparing Classification Results between N-ary and Binary Problems. In: Studies in Computational Intelligence, Springer-Verlag, volume 43. pp. 277–301.
7. Diri B, Albayrak S (2008) Visualization and analysis of classifiers performance in multi-class medical data. Expert Systems with Applications 34: 628–634.
8. Hanley J, McNeil B (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143: 29–36.
9. Bradley A (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30: 1145–1159.
10. Everson R, Fieldsend J (2006) Multi-class ROC analysis from a multi-objective optimisation perspective. Pattern Recognition Letters 27: 918–927.
11. Landgrebe T, Duin R (2005) On Neyman-Pearson optimisation for multiclass classifiers. In: Proceedings 16th Annual Symposium of the Pattern Recognition Association of South Africa. PRASA, pp. 165–170.
12. Landgrebe T, Duin R (2006) A simplified extension of the Area under the ROC to the multiclass domain. In: Proceedings 17th Annual Symposium of the Pattern Recognition Association of South Africa. PRASA, pp. 241–245.
13. Landgrebe T, Duin R (2008) Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis. IEEE Transactions Pattern Analysis Machine Intelligence 30: 810–822.
14. Ferri C, Hernández-Orallo J, Salido M (2003) Volume under the ROC surface for multi-class problems. In: Proceedings of 14th European Conference on Machine Learning. Springer-Verlag, pp. 108–120.
15. Van Calster B, Van Belle V, Condous G, Bourne T, Timmerman D, et al. (2008) Multi-class AUC metrics and weighted alternatives. In: Proceedings 2008 International Joint Conference on Neural Networks, IJCNN08. IEEE, pp. 1390–1396.
16. Li Y (2009) A generalization of AUC to an ordered multi-class diagnosis and application to lon-gitudinal data analysis on intellectual outcome in pediatric brain-tumor patients. Ph.D. thesis, College of Arts and Sciences, Georgia State University.
17. Hand D, Till R (2001) A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning 45: 171–186.
18. Freitas C, De Carvalho J, Oliveira J Jr, Aires S, Sabourin R (2007) Confusion matrix disagreement for multiple classifiers. In: Rueda L, Mery D, Kittler J, editors, Proceedings of 12th Iberoamerican Congress on Pattern Recognition, CIARP 2007, LNCS 4756. Springer-Verlag, pp. 387–396.
19. Freitas C, De Carvalho J, Oliveira J Jr, Aires S, Sabourin R (2007) Distance-based Disagreement Classifiers Combination. In: Proceedings of the International Joint Conference on Neural Networks, IJCNN 2007. IEEE, pp. 2729–2733.
20. Matthews B (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta - Protein Structure 405: 442–451.
21. Baldi P, Brunak S, Chauvin Y, Andersen C, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16: 412–424.
22. The MicroArray Quality Control (MAQC) Consortium (2010) The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models. Nature Biotechnology 28: 827–838.
23. Gorodkin J (2004) Comparing two K-category assignments by a K-category correlation coefficient. Computational Biology and Chemistry 28: 367–374.
24. Supper J, Spieth C, Zell A (2007) Reconstructing Linear Gene Regulatory Networks. In: Marchiori E, Moore J, Rajapakse J, editors, Proceedings of the 5th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, EvoBIO2007, LNCS 4447. Springer-Verlag, pp. 270–279.

25. Stokic D, Hanel R, Thurner S (2009) A fast and efficient gene-network reconstruction method from multiple over-expression experiments. BMC Bioinformatics 10: 253.
26. Shannon C (1948) A Mathematical Theory of Communication. The Bell System Technical Journal 27: 379–423, 623–656.
27. van Son R (1994) A method to quantify the error distribution in confusion matrices. Technical Report IFA Proceedings 18, Institute of Phonetic Sciences, University of Amsterdam.
28. Abramson N (1963) Information theory and coding. McGraw-Hill, 201 pp.
29. Sindhwani V, Bhattacharge P, Rakshit S (2001) Information theoretic feature crediting in multiclass Support Vector Machines. In: Grossman R, Kumar V, editors, Proceedings First SIAM International Conference on Data Mining, ICDM01. SIAM, pp. 1–18.
30. Wei JM, Yuan XJ, Hu QH, Wang SQ (2010) A novel measure for evaluating classifiers. Expert Systems with Applications 37: 3799–3809.
31. Wei JM, Yuan XJ, Yang T, Wang SQ (2010) Evaluating Classifiers by Confusion Entropy. Information Processing & Management Submitted.
32. Huang J, Ling C (2005) Using AUC and Accuracy in Evaluating Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering 17: 299–310.
33. Sørlie T, Tibshirani R, Parker J, Hastie T, Marron JS, et al. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences 100: 8418–8423.
34. Hastie T, Tibshirani R, Friedman JH (2003) The Elements of Statistical Learning. Springer.
35. Hand D (2010) Evaluating diagnostic tests: The area under the ROC curve and the balance of errors. Statistics in Medicine 29: 1502–1510.