Good Classification Measures and How to Find Them

Martijn Gösgens
Eindhoven University of Technology
Eindhoven, The Netherlands
research@martijngosgens.nl

Anton Zhiyanov
Yandex Research, HSE University
Moscow, Russia
zhiyanovap@gmail.com

Alexey Tikhonov
Yandex
Berlin, Germany
altsoph@gmail.com

Liudmila Prokhorenkova
Yandex Research, MIPT, HSE University
Moscow, Russia
ostroumova-la@yandex.ru

Abstract

Several performance measures can be used for evaluating classification results: accuracy, F-measure, and many others. Can we say that some of them are better than others, or, ideally, choose one measure that is best in all situations? To answer this question, we conduct a systematic analysis of classification performance measures: we formally define a list of desirable properties and theoretically analyze which measures satisfy which properties. We also prove an impossibility theorem: some desirable properties cannot be simultaneously satisfied. Finally, we propose a new family of measures satisfying all desirable properties except one. This family includes the Matthews correlation coefficient and a so-called symmetric balanced accuracy that was not previously used in classification literature. We believe that our systematic approach gives an important tool to practitioners for adequately evaluating classification results.

1 Introduction

Classification is a classic machine learning task that is used in countless applications. To evaluate classification results, one has to compare the predicted labeling of a given set of elements with the actual (true) labeling. For this, *performance measures* are used, and there are many well-known ones like accuracy, F-measure, and so on [12]. The fact that different measures behave differently is known throughout the literature [2, 12, 23]. For instance, accuracy is known to be biased towards the majority class. Thus, different measures may lead to different evaluation results, and it is important to choose an appropriate measure. While there are attempts to compare performance measures and describe their properties [3, 5, 12, 15, 21, 23], the problem still lacks a systematic approach, and our paper aims at filling this gap.[1] Our research is particularly motivated by a recent paper [11] providing a systematic analysis of evaluation measures for the *clustering* task. We transfer many proposed properties to the classification problem and extend the research by adding more properties, new measures, and novel theoretical results.

To provide a systematic comparison of performance measures, we formally define a list of properties that are desirable across various classification tasks. The proposed properties can be applied both to binary and multiclass problems. Some properties are intuitive and straightforward, like symmetry, while others are more tricky. A particularly important property is called *constant baseline*. It requires a measure not to be biased towards particular predicted class sizes. For each measure and each

[1] We describe related research in detail in Appendix A.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

property, we formally prove or disprove that the property is satisfied. We believe that this analysis is essential for better understanding the differences between the performance measures.

Then, we analyze relations between different properties in the binary case and prove an impossibility theorem: it is impossible for a performance measure to be linearly transformable to a metric and simultaneously have the constant baseline property. This means that at least one of these properties has to be discarded. If we relax the set of properties by discarding the distance requirement, the remaining ones can be simultaneously satisfied. In fact, we propose a family of measures called Generalized Means (GM), satisfying all the properties except distance and generalizing the well-known Matthews Correlation Coefficient (CC). In addition to CC, this class also contains another interesting measure that we name *Symmetric Balanced Accuracy*. To the best of our knowledge, this measure has not been previously used for classification evaluation.[2] If we instead discard the constant baseline (but keep its approximation), then the arccosine of CC is a measure satisfying all the properties.

We also demonstrate through a series of experiments that different performance measures can be inconsistent in various situations. We notice that measures having more desirable properties are usually more consistent with each other.

We hope that our research will motivate further studies analyzing the properties of performance measures for classification and other problems since there are still plenty of questions to be answered.

2 Performance measures for classification

In this section, we define measures that are commonly used for evaluating classification results. Classification problems can be divided into binary, multiclass, and multilabel. In this paper, we focus on *binary* and *multiclass* and leave multilabel for future research. There are several types of performance measures: *threshold* measures assume that predicted labels deterministically assign each element to a class (e.g., accuracy); *probability* measures assume that the predicted labels are soft and compare these probabilities with the actual outcome (e.g., the cross-entropy loss); *ranking* measures take into account the relative order of the predicted soft labels, i.e., quantify whether the elements belonging to a class have higher predicted probabilities compared to other elements (e.g., area under the ROC curve or average precision). Our research focuses on threshold measures.

Now we introduce notation needed to formally define binary and multiclass threshold measures.[3] Let $n > 0$ be the number of elements in the dataset and let $m \geq 2$ denote the number of classes. We assume that there is *true labeling* classifying elements into m classes and also *predicted labeling*. Let $\mathcal{C}$ be the confusion matrix: each matrix element c_{ij} denotes the number of elements with true label i and predicted label j. For binary classification, c_{11} is *true positive* (TP), c_{00} is *true negative* (TN), c_{10} is *false negative* (FN), and c_{01} is *false positive* (FP). We use the notation $a_i = \sum_{j=0}^{m-1} c_{ij}, b_i = \sum_{j=0}^{m-1} c_{ji}$ for the sizes of i-th class in the true and predicted labelings, respectively. Finally, we denote classification measures by $M(\mathcal{C})$ or $M(A, B)$, where A and B are true and predicted labelings, and write $M(c_{11}, c_{10}, c_{01}, c_{00})$ for binary ones.

Table 1 (above the line) lists several widely used classification measures. The most well-known is *Accuracy* which is the fraction of correctly classified elements. Accuracy is known to be biased towards the majority class, so it is not appropriate for unbalanced problems. To overcome this, *Balanced Accuracy* re-weights the terms to treat all classes equally. *Cohen's Kappa* uses a different approach to overcome this bias: it corrects the number of correctly classified samples by the expected value obtained by a random classifier [4]. *Matthews Correlation Coefficient* is the Pearson correlation coefficient between true and predicted labelings for binary classification [10]. For the multiclass case, covariance is computed for each class, and the obtained values are averaged before computing the correlation coefficient. Finally, *Confusion Entropy* computes the entropy of the misclassification distribution for each class and combines the obtained values, see Table 1 and [25] for the details.[4]

[2] For clustering evaluation, there is an analog known as *Sokal&Sneath's measure* [11].

[3] For convenience, we list the notation used in the paper in Table 7 in Appendix.

[4] There can be cases when a class is not present in the predicted labels. Then, some measures may contain division by zero. A proper way to fill in such singularities is discussed in Appendix B.

Table 1: Commonly used (above the line) and novel (below the line) validation measures

	Binary	Multiclass
F-measure (F_β)	$\frac{(1+\beta^2)\cdot c_{11}}{(1+\beta^2)\cdot c_{11}+\beta^2\cdot c_{10}+c_{01}}$	micro / macro / weighted
Jaccard (J)	$\frac{c_{11}}{c_{11}+c_{10}+c_{01}}$	micro / macro / weighted
Matthews Coefficient (CC)	$\frac{c_{11}c_{00}-c_{01}c_{10}}{\sqrt{b_1\cdot a_1\cdot b_0\cdot a_0}}$	$\frac{n\sum_{i=0}^{m-1} c_{ii}-\sum_{i=0}^{m-1} b_i a_i}{\sqrt{\left(n^2-\sum_{i=0}^{m-1} b_i^2\right)\left(n^2-\sum_{i=0}^{m-1} a_i^2\right)}}$
Accuracy (Acc)	$\frac{\sum_{i=0}^{m-1} c_{ii}}{n}$	
Balanced Accuracy (BA)	$\frac{1}{m}\sum_{i=0}^{m-1} \frac{c_{ii}}{a_i}$	
Cohen's Kappa (κ)	$\frac{\sum_{i=0}^{m-1} c_{ii}-\frac{1}{n}\sum_{i=0}^{m-1} a_i b_i}{n-\frac{1}{n}\sum_{i=0}^{m-1} a_i b_i}$	
Confusion Entropy (CE)	$-\frac{1}{2n}\sum_{i,j:i\neq j}\left(c_{ji}\log_{2m-2}\frac{c_{ji}}{a_j+b_j}+c_{ij}\log_{2m-2}\frac{c_{ij}}{a_j+b_j}\right)$	
Symmetric Balanced Accuracy (SBA)	$\frac{1}{2m}\sum_{i=0}^{m-1}\left(\frac{c_{ii}}{a_i}+\frac{c_{ii}}{b_i}\right)$	
Generalized Means (GM)	$\frac{n\,c_{11}-a_1 b_1}{\sqrt[r]{\frac{1}{2}\left(a_1^r a_0^r+b_1^r b_0^r\right)}}$	micro / macro / weighted
Correlation Distance (CD)	$\frac{1}{\pi}\arccos(\mathrm{CC})$	

Some measures are exclusively defined for binary classification. In this case, the classes are often referred to as 'positive' and 'negative'. *Jaccard* measures the fraction of correctly detected positive examples among all positive ones (both in true and predicted labelings). ***F-measure*** is the (possibly weighted) harmonic mean of Recall (c_{11}/a_1) and Precision (c_{11}/b_1). For measures that do not have a natural multiclass variant, there are several universal extensions obtained via *averaging* the results for m one-vs-all binary classifications [13]. For each such one-vs-all classification, a particular class i is considered positive while all other classes are grouped to a negative class.

Micro averaging sums up all binary confusion matrices corresponding to m one-vs-all classifications. Formally, it sets true positive as $\sum_{i=0}^{m-1} c_{ii}$, false negative and false positive as $n-\sum_{i=0}^{m-1} c_{ii}$, true negative as $(m-2)n+\sum_{i=0}^{m-1} c_{ii}$. Then, a given binary measure is applied to the obtained matrix.

Macro averaging computes the metric values for m binary classification sub-problems and then averages the results: $\frac{1}{m}\sum_{i=0}^{m-1} M(c_{ii}, a_i-c_{ii}, b_i-c_{ii}, n-a_i-b_i+c_{ii})$, where $M(\cdot)$ is a given binary measure. Note that macro averaging gives equal weights to all one-vs-all binary classifications.

In contrast, *weighted averaging* weights one-vs-all binary classifications according to the sizes of the corresponding classes: $\frac{1}{n}\sum_{i=0}^{m-1} a_i \cdot M(c_{ii}, a_i-c_{ii}, b_i-c_{ii}, n-a_i-b_i+c_{ii})$.

3 Properties of validation measures

As clearly seen from the above discussion, there are many options for a classification validation. In this section, we propose a formal approach that allows for a better understanding the differences between the measures and for making an informed decision among them for a particular application. For this, we propose properties of validation measures that can be useful across various applications and formally checks which measures satisfy which properties. In this regard, we follow an approach proposed in [11] for comparing validation measures for *clustering* tasks.

First, we observe that some theoretical results from [11] are related to *binary* classification measures. Indeed, a popular subclass of clustering validation measures are *pair-counting* ones. Such measures are defined in terms of the values $N_{11}, N_{10}, N_{01}, N_{00}$ that essentially define a confusion matrix for binary classification on *element pairs*. Thus, replacing N_{ij} in pair-counting clustering measures by c_{ij}, results in *binary* classification measures. We refer to Appendix B for the correspondence of some classification and clustering measures. In particular, Accuracy is equivalent to Rand, while Cohen's Kappa corresponds to Adjusted Rand. This equivalence allows us to transfer some of the results from [11] to the context of binary classification. However, an important contribution of our work is the extension of the properties and analysis to the multiclass case. We also prove an impossibility

Table 2: Properties of validation measures and averagings, ✓/✗ indicates that property is satisfied only in binary case

Measure	Max	Min	CSym	Sym	Dist	Mon	SMon	CB	ACB
F_1 (binary)	✓	✗	✗	✓	✗	✓	✗	✗	✗
J (binary)	✓	✗	✗	✓	✓	✓	✗	✗	✗
CC	✓	✓/✗	✓	✓	✗	✓	✓/✗	✓	✓
Acc	✓	✓	✓	✓	✓	✓	✓	✗	✗
BA	✓	✓	✓	✗	✗	✓	✓	✓	✓
κ	✓	✗	✓	✓	✗	✓	✗	✓	✓
CE	✓	✗	✓	✓	✗	✗	✗	✗	✗
SBA	✓	✓	✓	✓	✗	✓	✓	✓	✓
GM (binary)	✓	✓	✓	✓	✗	✓	✓	✓	✓
CD	✓	✓/✗	✓	✓	✓	✓	✓/✗	✗	✓
Preserving properties by various averaging types									
Micro	✓	✗	✓	✓	✓	✓	✗	✗	✗
Macro	✓	✗	✓	✓	✓	✓	✗	✓	✓
Weighted	✓	✗	✓	✗	✗	✓	✗	✓	✓

theorem stating that some of the desirable properties cannot be simulataneusly satisfied and develop a new family of measures having all properties except one.

Similarly to [11], we note that all the discussed properties are invariant under linear transformations and interchanging true and predicted labelings. Hence, we may restrict to measures for which higher values indicate higher similarity between classifications.

Table 2 summarizes our findings: for each measure, we mathematically prove or disprove each desirable property. Further in this section, we refer only to known measures (above the line), while the remaining ones will be defined and analyzed in Section 4. In addition to individual measures, we also analyze the properties of micro, macro, and weighted multiclass averagings: for each averaging, we analyze whether it preserves a given property, assuming the binary classification measure satisfies it. All the proofs can be found in Appendix C. Let us now define and motivate each property.

3.1 Minimal and maximal agreement

These properties make the upper and lower range of a performance measure interpretable. The *maximal agreement* property requires the measure to have an upper bound that is only achieved when the compared labelings are identical.

Definition 1. *We say that a measure M satisfies* maximal agreement *if there exists a constant $c_{\max}$ such that for all $\mathcal{C}$, $M(\mathcal{C}) \leq c_{\max}$ with equality iff $\mathcal{C}$ is diagonal.*

Also, for a given true labeling, there are several "worst" predictions, i.e., labelings that are wrong everywhere. This leads to the following property.

Definition 2. *We say that a measure M satisfies* minimal agreement *if there exists a constant $c_{\min}$ such that for all $\mathcal{C}$, $M(\mathcal{C}) \geq c_{\min}$ with equality iff the diagonal of $\mathcal{C}$ is zero, i.e., $c_{ii} = 0$ for all i.*

These properties allow for an easy and intuitive interpretation of the measure's values. While all of the measures in Table 2 do satisfy maximal agreement, there are popular measures such as Recall (c_{11}/a_1) and Precision (c_{11}/b_1) that do not satisfy this property as the maximum can also be achieved when the compared classifications are not identical. For minimal agreement, there are many performance measures that violate it. For example, Cohen's Kappa is obtained from Accuracy by subtracting the expected value of Accuracy and normalizing the result. As a result of the particular normalization used, it has minimal value $-\left(\sum_{i=0}^{m-1} a_i b_i\right) / \left(n^2 - \sum_{i=0}^{m-1} a_i b_i\right)$, which is clearly not constant.

If a binary measure satisfies maximum agreement, then its multiclass variant obtained via micro, macro, or weighted averaging also satisfies maximum agreement as each one-vs-all binary classification agrees maximally. This does not hold for minimal agreement: though each one-vs-all binary classification will have zero true positives, the number of true negatives may still be positive.

3.2 Symmetry

Definition 3. *We say that a measure M is* symmetric *if $M(\mathcal{C}) = M(\mathcal{C}^T)$ holds for all $\mathcal{C}$.*

In other words, we require symmetry with respect to interchanging predicted and true labels. This property is often desirable since similarity is usually understood as a symmetric concept. However, in some specific applications, there may be reasons to treat the true and predicted labelings differently and thus use an asymmetric measure. An example of an asymmetric measure is Balanced Accuracy.

Let us also introduce *class-symmetry*, i.e., invariance to permuting the classes.

Definition 4. *We say that a measure M is* class-symmetric *if, for any permutation π of the classes $\{1, \ldots, m\}$ and any confusion matrix $\mathcal{C}$, $M(\mathcal{C}) = M(\tilde{\mathcal{C}})$ holds, where $\tilde{\mathcal{C}}$ is given by $\tilde{c}_{ij} = c_{\pi(i),\pi(j)}$.*

Note that known multiclass measures are all class-symmetric, while in binary classification tasks, there can be asymmetry between 'positive' and 'negative' classes. Examples of well-known class-asymmetric binary classification measures are Jaccard and F_1.

3.3 Distance

In some applications, it is desirable to have a distance interpretation of a measure: whenever a labeling A is similar to B, while B is similar to C, it should intuitively hold that A is also somewhat similar to C. For instance, it can be the case that the *actual* labels are unknown, and the labeling A is only an approximation of the truth. Then, we would want the similarity between predicted labels and A to be not too different from the similarity between predicted and the actual true labels. This would be guaranteed if the measure is a distance.

Definition 5. *A measure has* distance *property if it can be linearly transformed to a metric distance.*

A function $d(A, B)$ is a metric distance if it is symmetric, nonnegative, equals zero only when $A = B$, and satisfies the triangle inequality $d(A, C) \leq d(A, B) + d(B, C)$. Note that the first requirement is equivalent to symmetry (Definition 3), while the second and third imply maximal agreement (Definition 1). Furthermore, note that if d is a distance, then $c \cdot d$ is also a distance for any $c > 0$. Therefore, we can conclude that M is a distance if and only if M satisfies symmetry and maximal agreement while $c_{\max} - M(A, B)$ satisfies the triangle inequality.

While most of the measures cannot be linearly transformed to a distance, some measures do satisfy this property. For example, the Jaccard measure can be transformed to the Jaccard distance $1 - \mathrm{J}(A, B)$. Similarly, Accuracy can be transformed to a distance by $1 - \mathrm{Acc}(A, B)$.

3.4 Monotonicity

Monotonicity is one of the most important properties of a similarity measure: intuitively, changing one labeling such that it becomes more similar to the other ought to result in an increase of the similarity score. Then, in order to formalize monotonicity, we need to determine what kind of changes make the classifications A and B more similar to each other. The simplest option is to take one element on which A and B disagree and resolve this disagreement.

Definition 6. *A measure M is* monotone *if $M(\mathcal{C}) < M(\tilde{\mathcal{C}})$ for any confusion matrices $\mathcal{C}$ and $\tilde{\mathcal{C}}$ where $\tilde{\mathcal{C}}$ is obtained from $\mathcal{C}$ by decrementing an off-diagonal entry c_{ab} and incrementing c_{aa} or c_{bb}.*

This definition defines a partial ordering over confusion matrices with the same total number of elements. However, we can relax the latter restriction and obtain the following, stronger notion of monotonicity that defines a partial ordering across different numbers of elements.

Definition 7. *A measure M is* strongly monotone *if $M(\mathcal{C}) < M(\tilde{\mathcal{C}})$ for any confusion matrix $\mathcal{C}$ that has at least one positive off-diagonal entry and every $\tilde{\mathcal{C}}$ that has at least one positive diagonal entry and is obtained from $\mathcal{C}$ by either increasing a diagonal entry or decreasing an off-diagonal entry.*

Note that we have to require that $\mathcal{C}$ is not a diagonal matrix, since otherwise it would contradict the maximal agreement property as $M(\mathcal{C}) = c_{\max} \geq M(\tilde{\mathcal{C}})$ holds when $\mathcal{C}$ is diagonal. Similarly, we must require $\tilde{\mathcal{C}}$ to not be zero on the whole diagonal in order to not contradict minimal agreement.

While almost all measures in Table 2 (except for CE) satisfy monotonicity from Definition 6, many fail to satisfy strong monotonicity. In particular, some widely used measures such as F_1, Jaccard and Cohen's Kappa do not satisfy this intuitive property.

3.5 Constant baseline

The constant baseline is perhaps the most important non-trivial property. On the one hand, it ensures that a measure is not biased towards labelings with particular class sizes $b_1, \ldots, b_m$. On the other hand, it also ensures some interpretability for 'mediocre' predictions.

Intuitively, if predicted labels are drawn at random and independently of the true labels, we would expect them to have a low similarity with the true labels. Then, if another prediction has a similarly low score, we can say that it is roughly as bad as a random guess. However, this is only possible when such random classifications achieve similar scores, independent from their class sizes. To formalize this, let $U(b_1, \ldots, b_m)$ denote the uniform distribution over labelings with class sizes $b_1, \ldots, b_m$. We say that the class sizes $b_1, \ldots, b_m$ are *unary* if $b_i = n$ for some $i \in \{1, \ldots, m\}$. That is, if all elements get classified to the same class so that $U(b_1, \ldots, b_m)$ is a constant distribution.

Definition 8. *We say that a measure M has a* constant baseline *property if there exists $c_{base}(m)$ that does not depend on n but may depend on m, such that for any true labeling A and non-unary class sizes $b_1, \ldots, b_m$, it holds that $\mathbb{E}_{B \sim U(b_1,\ldots,b_m)}[M(A, B)] = c_{base}(m)$.*

Note that we need to require the class sizes to be non-unary: if these class sizes are unary, we will have contradictions with maximal and minimal agreement when the class sizes of A are also unary. Many popular measures such as F_1, Accuracy, and Jaccard do not have a constant baseline. Furthermore, some measures that do have a constant baseline were deliberately designed to have one. For example, Cohen's Kappa was obtained from Accuracy by correcting it for chance. While our definition of the constant baseline does allow for a baseline $c_{\text{base}}(m)$ that depends on the number of classes m, some measures such as the Matthews Coefficient and Cohen's Kappa have a baseline that is constant w.r.t. m.

All of the measures that satisfy constant baseline turn out to be linear functions in terms of c_{ii} for fixed class sizes $a_1, \ldots, a_m$ and $b_1, \ldots, b_m$. For such measures, linearity of the expectation can be utilized to easily compute the baseline by substituting the expected values $\mathbb{E}_{B \sim U(b_1,\ldots,b_m)}[c_{ii}] = \frac{a_i b_i}{n}$. Thus, we also propose the following relaxation of the constant baseline property.

Definition 9. *A measure M is said to have an* approximate constant baseline *if there exists a function $c_{base}(m)$ that does not depend on n but may depend on m such that for any class sizes $a_1, \ldots, a_m$ and any non-unary $b_1, \ldots, b_m$, $M(\bar{\mathcal{C}}) = c_{base}(m)$, where $\bar{c}_{ij} = \frac{a_i b_j}{n}$.*

The advantage of this relaxation is that it allows us to non-linearly transform measures while still maintaining an approximate constant baseline. Take for example Matthews Coefficient: it cannot be linearly transformed to a distance while the transformations $\text{CD}(A, B) = \frac{1}{\pi} \arccos(\text{CC}(A, B))$ and $\sqrt{2(1 - \text{CC}(A, B))}$ do yield distances. Because Matthews Coefficient has a constant baseline, these non-linear transformations have an approximate constant baseline, see Section 4 for more details.

As can be seen from Table 2, there is no measure satisfying all the properties. In particular, there is no measure having both distance and constant baseline. In the next section, we show why this is not a coincidence.

4 An impossibility theorem for classification

In this section, we focus on binary classification and more deeply analyze the relations between the properties discussed above. Unfortunately, it turns out that the properties introduced in the previous section cannot all be satisfied simultaneously.

Theorem 1. *There is no binary classification measure that simultaneously satisfies the monotonicity, distance, and constant baseline properties.*

Proof. Let A be a labeling with a single positive and $n - 1$ negatives. Let B_1 be a random labeling with a single positive and let B_2 be a random labeling with two positives. The constant baseline

requires $\mathbb{E}[M(A, B_1)] = \mathbb{E}[M(A, B_2)]$, which gives

$$\frac{1}{n}c_{\max} + \frac{n-1}{n}M(0,1,1,n-2) = \frac{2}{n}M(1,0,1,n-2) + \frac{n-2}{n}M(0,1,2,n-3),$$

which we rewrite to

$$2M(1,0,1,n-2) - c_{\max} = (n-1)M(0,1,1,n-2) - (n-2)M(0,1,2,n-3). \tag{1}$$

Now, we consider a labeling C with a single positive that does not coincide with the positive of A and a labeling B that has two positives which are the positives of A and C. The triangle inequality tells us that

$$c_{\max} - M(0,1,1,n-2) \leq 2c_{\max} - M(1,1,0,n-2) - M(1,0,1,n-2) = 2(c_{\max} - M(1,1,0,n-2)),$$

where the last step follows from symmetry (implied by distance). This is rewritten to

$$2M(1,1,0,n-2) - c_{\max} \leq M(0,1,1,n-2). \tag{2}$$

Combining (1) and (2), we obtain

$$(n-1)M(0,1,1,n-2) - (n-2)M(0,1,2,n-3) \leq M(0,1,1,n-2).$$

We rewrite this to $M(0,1,1,n-2) \leq M(0,1,2,n-3)$, which clearly contradicts monotonicity. □

Thus, we have to discard one of these properties. Obviously, discarding monotonicity would be highly undesirable, since higher values would then not necessarily indicate higher similarity. For this reason, we analyze what happens if we discard either *distance* or *constant baseline*. All the results stated below are proven in Appendix D.

Discarding distance Assuming some additional smoothness conditions that are, however, satisfied by all measures discussed in this paper, we prove the following result.

Theorem 2. *All binary measures that satisfy all properties except distance must be of the form*

$$s\left(\frac{a_0 a_1}{n^2}, \frac{b_0 b_1}{n^2}\right) \cdot \frac{n c_{11} - a_1 b_1}{n^2},$$

where the normalization factor $s(a,b)$ needs to satisfy some additional properties listed in Theorem 3.

This class of measures is quite wide and contains many unelegant measures. An interesting subclass can be obtained if we normalize by the generalized mean, i.e., take $s(a,b)^{-1} = (\frac{1}{2}a^r + \frac{1}{2}b^r)^{1/r}$.

Definition 10. *For $r \in \mathbb{R}$, we define* Generalized Means *measures as*

$$\mathrm{GM}_r = \frac{n\, c_{11} - a_1 b_1}{\sqrt[r]{\frac{1}{2}\left(a_1^r a_0^r + b_1^r b_0^r\right)}}\,.$$

Statement 1. *For any $r \in \mathbb{R}$, the measure GM_r satisfies all properties except for being a distance.*

Let us show that the generalized means measures contain two interesting special cases.

Statement 2. *If $r \to 0$ (corresponding to the geometric mean),* $\mathrm{GM}_r(\mathcal{C}) \to \mathrm{CC}(\mathcal{C})$.
If $r = -1$ (corresponding to the harmonic mean), $\mathrm{GM}_{-1}(\mathcal{C}) = \mathrm{BA}(\mathcal{C}) + \mathrm{BA}(\mathcal{C}^\top) - 1$.

Thus, for $r = -1$ generalized means is equivalent to the measure $\frac{1}{2}\left(\mathrm{BA}(\mathcal{C}) + \mathrm{BA}(\mathcal{C}^\top)\right)$ that we call *Symmetric Balanced Accuracy* (SBA). To the best of our knowledge, this measure has not been used in the classification literature. However, in the clustering literature, a similar measure is known as Sokal&Sneath's measure [1, 11]. Interestingly, SBA preserves its properties for the multiclass case.

Statement 3. *SBA satisfies all properties except for being a distance for any $m \geq 2$.*

Discarding (exact) constant baseline Note that Theorem 1 only proves an impossibility for the *exact* constant baseline, but not the *approximate* constant baseline.

Statement 4. *The measures* $\mathrm{CD}(A,B) := \frac{1}{\pi}\arccos(\mathrm{CC}(A,B))$ *and* $\mathrm{CD}'(A,B) := \sqrt{2(1-\mathrm{CC}(A,B))}$ *satisfy all properties except the exact constant baseline, but including the approximate constant baseline.*

Following [11], we call the measure $\frac{1}{\pi}\arccos(\mathrm{CC}(A,B))$ *correlation distance* (CD). In Appendix D, we prove the following (see Appendix D.2 for the details).

Statement 5. *CD approximates a constant baseline with one order of precision better than CD′.*

Essentially, this is a consequence of the fact that the transformation $\frac{1}{\pi}\arccos(\mathrm{CC})$ is a symmetric function around the constant baseline $\mathrm{CC}=0$ while $\sqrt{2(1-\mathrm{CC})}$ is not. In more detail, we show that the leading error term of CD′ is of the order $\mathbb{E}[\mathrm{CC}(A,B)^2]$ while the leading error term for CD is of the order $\mathbb{E}[\mathrm{CC}(A,B)^3]$. Currently, we are not aware of other distance measures for which the constant baseline is approximated up to the same order of precision as CD. We thus argue that for binary classification tasks where a distance interpretation is desirable, the Correlation Distance is the most suitable measure.

5 Inconsistency of measures in practice

In this section, we conduct several experiments that demonstrate how often performance measures may disagree in practice in different scenarios. These experiments demonstrate the importance of the problem considered in this paper and also show which pairs of measures are usually more consistent than others. For binary classification, we consider all measures from Table 1. For the F-measure, we take $\beta=1$, for Generalized Means, we consider $r=1$. Recall that SBA and CC are also instances of GM with $r=-1$ and $r\to 0$, respectively. Furthermore, Jaccard is a monotone transformation of F_1, and CD is a monotone transformation of CC. Therefore, we omit CD and Jaccard from all inconsistency tables. The code for our experiments can be found on GitHub.[5]

5.1 Binary measures

Distinguishing measures for small datasets First, we construct simple examples showing the inconsistency of all pairs of binary classification measures. We say that two measures M_1 and M_2 are consistent on a triplet of classifications (A,B_1,B_2) if $M_1(A,B_1) * M_1(A,B_2)$ implies $M_2(A,B_1) * M_2(A,B_2)$, where $* \in \{>,<,=\}$. Otherwise, we say that the measures are inconsistent. We took $n\in\{2,3,\dots,10\}$ and went through all the possible triplets (A,B_1,B_2) of binary labelings of n elements (we additionally require that all labelings contain both classes). For each triplet, we check which pairs of measures are inconsistent. We say that a pair of measures is indistinguishable for a given n if it is consistent on all triplets.

Table 3 lists all measures that are indistinguishable for a given n. For $n=2$, all measures are always consistent. For $n=4$, we can distinguish Acc, F_1, and CE from other measures and each other. Interestingly, the remaining measures are those having the constant baseline property. Importantly, the most consistent measures are CC, SBA, and GM_1; and these are measures having the best properties according to our analysis. This supports our intuition that "good" measures agree with each other better than those having fewer desired properties. Additionally, in Appendix E.1, we list six triplets (A,B_1,B_2) with $n=10$ that discriminate all pairs of different measures.

Table 3: Indistinguishable measures

n	measures
2	[Acc, BA, F_1, κ, CE, GM_1, CC, SBA]
3	[Acc, BA, κ, GM_1, CC, SBA]
4-5	[BA, κ, GM_1, CC, SBA]
6-7	[GM_1, CC, SBA]
8	[CC, SBA]
9-10	—

Experiment within a weather forecasting service In this experiment, we aim at understanding whether the differences between measures may affect the decisions made while designing real systems. For this purpose, we conduct an experiment within the *Yandex.Weather* service.

[5] https://github.com/yandex-research/classification-measures

There is a model that predicts the presence/absence of precipitation at a particular location [14]. The prediction is made for 12 prediction intervals (*horizons*): from ten minutes to two hours. The original model returns the probability of precipitation, which can be converted to binary labels via a threshold. There are six thresholds used in this experiment, which lead to six different models. The measures were logged for 12 days. To sum up, for each threshold (model), each day, and each horizon, we have a confusion matrix that can be used to compute a performance measure.

For each pair of measures, we compute how often they are inconsistent according to the definition above. For this, we aggregate the results over all days and horizons. Table 4 shows that there are pairs of measures with substantial disagreement: e.g., accuracy and balance accuracy almost always disagree. This can be explained by the fact that accuracy has a bias towards the majority class, so it prefers a higher threshold, while balanced accuracy weighs true positives more heavily, so it prefers a lower threshold. In contrast, GM_1, CC, κ, and F_1 agree with each other much better. In Appendix E.1 we conduct a more detailed analysis. In particular, we separately consider the ten-minute and two-hour prediction horizons and show that behavior and consistency of measures significantly depend on the horizon as it defines the balance between true positives, true negatives, false positives, and false negatives. We also observe that CC and SBA perfectly agree for the ten-minute horizon but have noticeable disagreement for two hours.

Table 4: Inconsistency of binary measures for rain prediction, %

	Acc	BA	F_1	κ	CE	GM_1	CC	SBA
Acc	—	96.5	41.0	37.5	3.1	38.7	44.3	55.9
BA	96.5	—	55.6	58.9	99.7	57.7	52.0	40.4
F_1	41.0	55.6	—	3.3	44.2	2.2	3.4	15.0
κ	37.5	58.9	3.3	—	40.7	1.1	6.7	18.3
CE	3.1	99.7	44.2	40.7	—	41.9	47.5	59.1
GM_1	38.7	57.7	2.2	1.1	41.9	—	5.5	17.1
CC	44.3	52.0	3.4	6.7	47.5	5.5	—	11.4
SBA	55.9	40.4	15.0	18.3	59.1	17.1	11.4	—

5.2 Multiclass measures

In this section, we analyze multiclass measures. For all measures that are defined for the multiclass problems, we consider their standard expressions (if not stated otherwise). For other measures (F_1, Jaccard, GM_1), we use the macro averaging.

Image classification We conduct an experiment on ImageNet [19], a classic dataset for image classification. For this, we take the top-10 algorithms that are considered to be state-of-the-art at the moment of submission.[6] We check whether the leaderboard based on accuracy is consistent with the leaderboards based on other measures. Thus, we apply the models to the test set, compute the confusion matrices, and compare all measures defined in Table 1.

Notably, the ImageNet dataset is balanced. This makes all measures more similar to each other. For instance, accuracy and BA are equal in this scenario. Also, the *constant baseline* property discussed in Section 3.5 is especially important for *unbalanced* datasets. Thus, measures are *more consistent* on balanced data. Nevertheless, we notice that the ranking can be inconsistent starting from the algorithm ranked fifth on the leaderboard.

The (partial) results are shown in Table 5. Here we compare EfficientNet-B7 NoisyStudent [26] and Swin-B Transformer (patch size 4x4, window size 12x12, image size 384^2) [16] that are the fifth and sixth models in the leaderboard. One can see that the measures inconsistently rank the algorithms: confusion entropy, Jaccard, and SBA disagree with accuracy and other measures. Interestingly, while Jaccard and F_1 always agree for binary problems, they may disagree after the macro averaging, as we see in this case. Also, for one measure, different multiclass extensions may be inconsistent, as we see with macro averaging versus the standard definition of the multiclass correlation coefficient. More detailed results can be found in Appendix E.2.

Sentiment analysis In the previous experiment, we noticed that despite several disagreements, the measures usually rank the algorithms similarly. This can be caused by the fact that the test set of ImageNet is balanced: all classes have equal sizes. However, in practical applications, we

[6] https://github.com/rwightman/pytorch-image-models/blob/master/results/results-imagenet.csv (May 8, 2021).

Table 5: Inconsistent results on ImageNet, % (fifth and sixth models in the leaderboard)

	Acc/BA	F_1	J	κ	1−CE	GM_1	CC	CC^{macro}	SBA
Efficientnet	**86.46**	**86.30**	77.525	**86.44**	93.41	**86.28**	**86.44**	86.419	86.57
Swin	86.43	86.27	**77.531**	86.42	**93.51**	86.26	86.42	**86.423**	**86.61**

Table 6: Ranking algorithms according to different measures on SST-5: from 1 (best) to 7 (worst)

	Acc	BA	F_1	J	κ	CE	GM_1	CC	CC^{macro}	SBA
Flair+ELMo	1	1	1	1	1	1	1	1	1	1
Flair+BERT	2	4	5	5	4	2	5	2	2	2
SVM	3	3	3	3	3	5	3	3	4	4
Logistic	4	5	4	4	5	3	4	5	5	3
FastText	5	2	2	2	2	6	2	4	3	5
VADER	6	6	6	6	6	7	6	6	6	7
TextBlob	7	7	7	7	7	4	7	7	7	6

rarely encounter balanced data. Thus, we also consider an unbalanced classification task. In this experiment, we take the 5-class Stanford Sentiment Treebank (SST-5) dataset [22]. We compare the following algorithms: TextBlob, VADER, Logistic Regression, SVM, FastText, Flair+ELMo, and Flair+BERT [18]. Table 6 shows that different metrics rank the algorithms differently. Among the measures shown in the table, the only consistent rankings are the one provided by κ and BA and the second given by F_1, GM, and Jaccard. Note that the latter ranking significantly disagrees with the ranking by accuracy.

Appendix E.2 contains an additional experiment with an unbalanced multiclass dataset, where we show the inconsistency rates of the considered measures and different multiclass extensions.

6 Conclusion and future work

In this paper, we propose a systematic approach to analyzing classification performance measures: we propose several desirable properties and theoretically check each property for a list of measures. We also prove an impossibility theorem: some desirable properties cannot be simultaneously satisfied, so either distance or *exact* constant baseline has to be discarded.

Based on the properties we analyzed in this paper, we come to the following practical suggestions. If the distance requirement is needed, correlation distance seems to be the best option: it satisfies all the properties except for the exact constant baseline, which is still approximately satisfied. Otherwise, we suggest using one of generalized means, including correlation coefficient and symmetric balanced accuracy — they satisfy all the properties except distance. For binary classification, CC is a natural choice as it can be non-linearly transformed to a distance. For multiclass problems, symmetric balanced accuracy has an additional advantage: among the considered measures, only this one preserves its good properties in the multiclass case. Finally, we do not advise using averagings, but if needed, then macro averaging preserves more properties.

There are still many open questions and promising directions for future research. First, we would like to see whether one could construct a set of desirable properties that can be used as axioms to uniquely define one good measure (or a parametrized group of measures). Secondly, it is an open problem whether generalized means measures in general (or SBA in particular) can be converted to a distance via a continuous transformation. Finally, our work does not cover ranking and probability-based measures. Thus, we leave aside such widely used measures as cross-entropy and AUC. Formalizing and analyzing their properties is an important direction for future research.

Acknowledgments and Disclosure of Funding

Part of this work was done while Martijn Gösgens was visiting Yandex and Moscow Institute of Physics and Technology (Russia). The work of Martijn Gösgens is supported by the Netherlands Organisation for Scientific Research (NWO) through the Gravitation NETWORKS grant no. 024.002.003. The work of Liudmila Prokhorenkova is partially supported by the Ministry of Education and Science of the Russian Federation in the framework of MegaGrant 075-15-2019-1926 and by the Russian President grant supporting leading scientific schools of the Russian Federation NSh2540.2020.1.

The authors would like to thank Alexander Ganshin, Pert Vytovtov, and Eugenia Elistratova for providing the weather forecasting data.

References

[1] A. N. Albatineh, M. Niewiadomska-Bugaj, and D. Mihalko. On similarity indices and correction for chance agreement. *Journal of Classification*, 23(2):301–313, 2006.

[2] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010.

[3] D. Chicco and G. Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

[4] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[5] C. Cortes and M. Mohri. Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, 16(16):313–320, 2004.

[6] R. Delgado and J. D. Núñez-González. Enhancing confusion entropy (cen) for binary and multiclass classification. *PloS one*, 14(1):e0210264, 2019.

[7] R. Delgado and X.-A. Tibau. Why cohen's kappa should be avoided as performance measure in classification. *PloS one*, 14(9):e0222916, 2019.

[8] D. Dua and C. Graff. UCI machine learning repository, 2017.

[9] L. Flight and S. A. Julious. The disagreeable behaviour of the kappa statistic. *Pharmaceutical statistics*, 14(1):74–78, 2015.

[10] J. Gorodkin. Comparing two k-category assignments by a k-category correlation coefficient. *Computational biology and chemistry*, 28(5-6):367–374, 2004.

[11] M. Gösgens, L. Prokhorenkova, and A. Tikhonov. Systematic analysis of cluster similarity indices: How to validate validation measures. *International Conference on Machine Learning (ICML)*, 2021.

[12] M. Hossin and M. Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

[13] O. Koyejo, N. Natarajan, P. Ravikumar, and I. S. Dhillon. Consistent multilabel classification. In *NIPS*, volume 29, pages 3321–3329, 2015.

[14] V. Lebedev, V. Ivashkin, I. Rudenko, A. Ganshin, A. Molchanov, S. Ovcharenko, R. Grokhovetskiy, I. Bushmarinov, and D. Solomentsev. Precipitation nowcasting with satellite imagery. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2680–2688, 2019.

[15] C. X. Ling, J. Huang, H. Zhang, et al. Auc: a statistically consistent and more discriminating measure than accuracy. In *Ijcai*, volume 3, pages 519–524, 2003.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[17] D. M. W. Powers. The problem with kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 345–355, 2012.

[18] P. Rao. Fine grained sentiment classification. `https://github.com/prrao87/fine-grained-sentiment`, 2021.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[20] Scikit-learn. Clustering algorithms. `https://scikit-learn.org/stable/supervised_learning.html`, 2021.

[21] F. Sebastiani. An axiomatically derived measure for the evaluation of classification algorithms. In *Proceedings of the 2015 international conference on the theory of information retrieval*, pages 11–20, 2015.

[22] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

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

[24] V. Van Asch. Macro-and micro-averaged evaluation measures. *Technical report*, 2013.

[25] J.-M. Wei, X.-J. Yuan, Q.-H. Hu, and S.-Q. Wang. A novel measure for evaluating classifiers. *Expert Systems with Applications*, 37(5):3799–3809, 2010.

[26] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work may help towards reducing certain biases in research. For instance, asymmetric measures such as Jaccard may have a bias towards a specific class. If we would select a classifier based on such an asymmetric measure, it may result in outcomes that are unfair towards some classes. A similar problem occurs for measures with a bias towards the majority class (e.g., Accuracy). Thus, the bias towards the majority class could be even amplified with the poor metric selection. Our work could provide some clues on how to avoid such a situation.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sections 3 and 4.
   (b) Did you include complete proofs of all theoretical results? [Yes] The proofs are given in Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide link to the code.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5 and the supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We do not have any heavy computations and do not measure computation time.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 5 and the supplemental material.
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]