Noname manuscript No.
(will be inserted by the editor)

Evaluation Measures for Quantification: An Axiomatic Approach

Fabrizio Sebastiani

Received: July 20, 2020/ Accepted: July 20, 2020

Abstract Quantification is the task of estimating, given a set σ of unlabelled items and a set of classes $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$, the prevalence (or "relative frequency") in σ of each class $c_i \in \mathcal{C}$. While quantification may in principle be solved by classifying each item in σ and counting how many such items have been labelled with c_i, it has long been shown that this "classify and count" (CC) method yields suboptimal quantification accuracy. As a result, quantification is no longer considered a mere byproduct of classification, and has evolved as a task of its own. While the scientific community has devoted a lot of attention to devising more accurate quantification methods, it has not devoted much to discussing what properties an *evaluation measure for quantification* (EMQ) should enjoy, and which EMQs should be adopted as a result. This paper lays down a number of interesting properties that an EMQ may or may not enjoy, discusses if (and when) each of these properties is desirable, surveys the EMQs that have been used so far, and discusses whether they enjoy or not the above properties. As a result of this investigation, some of the EMQs that have been used in the literature turn out to be severely unfit, while others emerge as closer to what the quantification community actually needs. However, a significant result is that no existing EMQ satisfies all the properties identified as desirable, thus indicating that more research is needed in order to identify (or synthesize) a truly adequate EMQ.

1 Introduction

Quantification (also known as "supervised prevalence estimation" (Barranquero et al. 2013), or "class prior estimation" (du Plessis et al. 2017)) is

Fabrizio Sebastiani
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, Italy
E-mail: fabrizio.sebastiani@isti.cnr.it

the task of estimating, given a set σ of unlabelled items and a set of classes $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$, the relative frequency (or "prevalence") $p(c_i)$ of each class $c_i \in \mathcal{C}$, i.e., the fraction of items in σ that belong to c_i. When each item belongs to exactly one class, since $0 \leq p(c_i) \leq 1$ and $\sum_{c_i \in \mathcal{C}} p(c_i) = 1$, p is a *distribution* of the items in σ across the classes in $\mathcal{C}$ (the *true distribution*), and quantification thus amounts to estimating p (i.e., to computing a *predicted distribution* $\hat{p}$).

Quantification is important in many disciplines (such as e.g., market research, political science, the social sciences, and epidemiology) which usually deal with aggregate (as opposed to individual) data. In these contexts, classifying individual unlabelled instances is usually not a primary goal, while estimating the prevalence of the classes of interest in the data is. For instance, when classifying the tweets about a certain entity (e.g., a political candidate) as displaying either a Positive or a Negative stance towards the entity, we are usually not much interested in the class of a specific tweet: instead, we usually want to know the fraction of these tweets that belong to the class (Gao and Sebastiani 2016).

Quantification may in principle be solved via classification, i.e., by classifying each item in σ and counting, for all $c_i \in \mathcal{C}$, how many such items have been labelled with c_i. However, it has been shown in a multitude of works (see e.g., (Barranquero et al. 2015; Bella et al. 2010; Esuli and Sebastiani 2015; Forman 2008; Gao and Sebastiani 2016; Hopkins and King 2010)) that this "classify and count" (CC) method yields suboptimal quantification accuracy. Simply put, the reason of this suboptimality is that most classifiers are optimized for classification accuracy, and not for quantification accuracy. These two notions do not coincide, since the former is, by and large, inversely proportional to the sum $(FP_i + FN_i)$ of the false positives and the false negatives for c_i in the contingency table, while the latter is, by and large, inversely proportional to the absolute difference $|FP_i - FN_i|$ of the two. As a result, quantification has come to be no longer considered a mere byproduct of classification, and has evolved as a task of its own, devoted to designing methods and algorithms that deliver better prevalence estimates than CC (see (González et al. 2017) for a survey of methods and results).

While the scientific community working on quantification has devoted a lot of attention to devising new and more accurate quantification methods, it has not devoted much to discussing how quantification accuracy should be measured, i.e., what properties an *evaluation measure for quantification* (EMQ) should enjoy, and which EMQs should be adopted as a result. In experimental computer science, the properties of the evaluation measure that one uses are fundamental in order to ensure a correct comparison among systems, i.e., ensure that this comparison rewards the systems that deliver the most desirable results; these properties formalize what "desirable" actually means. In the quantification literature, sometimes new EMQs have been introduced without arguing why they are supposedly better than existing ones. As a result, there is no consensus (and, what is worse: no debate) in the field as to which EMQ (if any) is the best. Different authors use different EMQs without properly jus-

tifying their choice, and the consequence is that different results, even when obtained on the same dataset, are not comparable. Even worse, it may be the case that an improvement, sanctioned by an "inappropriate" EMQ, obtained by a newly proposed method with respect to a baseline, may correspond to no real improvement when measured according to an "appropriate" EMQ.

This paper attempts to shed some light on the issue of which evaluation measure(s) should be used for quantification. In order to do so, we (a) lie down a number of interesting properties that an EMQ may or may not enjoy, (b) discuss whether (or when) each of these properties is desirable, (c) survey the EMQs that have been used so far, and (d) discuss whether they enjoy or not the above properties. As a result of this investigation, some of the EMQs that have been used in the literature turn out to be severely unfit, while others emerge as closer to "what the quantification community actually needs". However, a significant result is that no existing measure satisfies all the properties identified as desirable, thus indicating that more research is needed in order to identify (or synthesize) a truly adequate EMQ.

This paper follows in the tradition of the so-called "axiomatic" approach to "evaluating evaluation" in information retrieval (see e.g., (Amigó et al. 2011; Busin and Mizzaro 2013; Ferrante et al. 2015, 2018; Moffat 2013; Sebastiani 2015)), which is based on describing (and often: arguing in favour of) a number of properties (that most of this literature calls – perhaps improperly – "axioms") that an evaluation measure for the task being considered should intuitively satisfy. The benefit of this approach is that it shifts the discussion from the evaluation measures to their properties, which amounts to shifting the discussion from a complex construction to its building blocks: once the scientific community has agreed on a set of properties (the building blocks), it then follows whether a given measure (the construction) is satisfactory or not.

The paper is structured as follows. In Section 2 we set the stage and define the scope of our investigation. In Section 3 we formally discuss properties that may or may not characterize an EMQ, and argue if and when it is desirable that an EMQ enjoys them. In Section 4 we turn to examining the actual measures that have been proposed or used in the quantification literature, and discuss whether they comply or not with the properties introduced in Section 3. Section 5 critically reexamines the results of Section 4, while Section 6 concludes, discussing aspects that the present work still leaves open and avenues for further research.

2 Evaluating Single-Label Quantification

Let us fix some notation. Symbols σ, σ', σ'', ... will each denote a *sample*, i.e., a nonempty set of unlabelled items, while symbols $\mathcal{C}$, $\mathcal{C}'$, $\mathcal{C}''$, ... will each denote a nonempty set of classes (or *codeframe*) across which the unlabelled items in a sample are distributed. Symbols c, c_1, c_2, ... will each denote an individual class. Given a class c_i, we will denote by σ_i the set of items in σ that belong to c_i; we will also denote by $|\sigma|$, $|\sigma'|$, $|\sigma''|$, ... the number of items

contained in samples σ, σ', σ'', ... Symbols p, p', p'' ..., will each denote a *true distribution* of the unlabelled items (either on the same sample σ or on different samples) across a codeframe $\mathcal{C}$, while symbols $\hat{p}$, $\hat{p}'$, $\hat{p}''$, ... will each denote a *predicted distribution* (or *estimator*), i.e., the result of estimating a true distribution;[1] symbol $\mathcal{P}$ will denote the (infinite) set of all distributions on $\mathcal{C}$.[2] Finally, symbols D, D', D'', ... will each denote an EMQ, while symbols π, π', π'', ... will denote properties that an EMQ may enjoy or not.

Similarly to classification, there are different quantification problems of applicative interest, based (a) on how many classes codeframe $\mathcal{C}$ contains, and (b) how many of the classes in $\mathcal{C}$ can be legitimately attributed to the same item. We characterize quantification problems as follows:

1. *Single-label quantification* (SLQ) is defined as quantification when each item belongs to exactly one of the classes in $\mathcal{C} = \{c_1, ..., c_{|\mathcal{C}|}\}$.
2. *Multi-label quantification* (MLQ) is defined as quantification when the same item may belong to any number of classes (zero, one, or several) in $\mathcal{C} = \{c_1, ..., c_{|\mathcal{C}|}\}$.
3. *Binary quantification* (BQ) may alternatively be defined
   (a) as SLQ with $|\mathcal{C}| = 2$ (in this case $\mathcal{C} = \{c_1, c_2\}$ and each item must belong to either c_1 or c_2), or
   (b) as MLQ with $|\mathcal{C}| = 1$ (in this case $\mathcal{C} = \{c\}$ and each item either belongs or does not belong to c).

Since BQ is a special case of SLQ (see bullet 3a above), any evaluation measure for SLQ is also an evaluation measure for BQ. Likewise, any evaluation measure for BQ is also an evaluation measure for MLQ, since evaluating a multi-label *quantifier* (i.e., a software artifact that estimates class prevalences) can be done by evaluating $|\mathcal{C}|$ binary quantifiers, one for each $c_i \in \mathcal{C}$. As a consequence, in this paper we focus on the evaluation of SLQ, knowing that all the solutions we discuss for SLQ also apply to BQ and MLQ.[3]

As already discussed, given a sample σ of items (single-)labelled according to $\mathcal{C} = \{c_1, ..., c_{|\mathcal{C}|}\}$, quantification has to do with determining, for each

[1] Consistently with most mathematical literature, we use the caret symbol (ˆ) to indicate estimation.

[2] In order to keep things simple we avoid overspecifying the notation, thus leaving some aspects of it implicit; e.g., in order to indicate a true distribution p of the unlabelled items in a sample σ across a codeframe $\mathcal{C}$ we will simply write p instead of the more cumbersome $p^{\mathcal{C}}_{\sigma}$, thus letting σ and $\mathcal{C}$ be inferred from context.

[3] In this paper we do not discuss the evaluation of *ordinal quantification* (OQ), defined as SLQ with a codeframe $\mathcal{C} = \{c_1, ..., c_{|\mathcal{C}|}\}$ on which a total order $c_1 \prec ... \prec c_{|\mathcal{C}|}$ is defined. Aside from reasons of space, the reasons for disregarding OQ is that there has been very little work on it (the only papers we know being (Da San Martino et al. 2016a,b; Esuli 2016)), and that only one measure for OQ (the *Earth Mover's Distance* – see (Esuli and Sebastiani 2010)) has been proposed and used so far. For the same reasons we do not discuss *regression quantification* (RQ), the task that stands to metric regression as single-label quantification stands to single-label classification. RQ has been studied even less than OQ, the only work appeared on this theme so far being, to the best of our knowledge, (Bella et al. 2014), which as an evaluation measure has proposed the *Cramér-von-Mises u-statistic* (see (Bella et al. 2014) for details).

$c_i \in \mathcal{C}$, the fraction $|\sigma_i|/|\sigma|$ of items in σ that are labelled by c_i. These $|\mathcal{C}|$ fractions actually form a distribution p of the items in σ across the classes in $\mathcal{C}$; quantification may thus be seen as generating a predicted distribution $\hat{p}(c)$ over $\mathcal{C}$ that approximates a true distribution $p(c)$ over $\mathcal{C}$. Evaluating quantification thus means measuring how well $\hat{p}(c)$ fits $p(c)$. We will thus be concerned with discussing the properties that a function that attempts to measure this goodness-of-fit should enjoy; we hereafter use the notation $D(p, \hat{p})$ to indicate such a function.[4]

In this paper we assume that the EMQs we are concerned with are measures of quantification error, and not of quantification accuracy. The reason for this is that most, if not all, the EMQs that have been used so far are indeed measures of error, so it would be slightly unnatural to discuss our properties with reference to quantification accuracy. Since any measure of accuracy can be turned into a measure of error (typically: by taking its negation), this is an inessential factor anyway.

3 Properties for SLQ Error Measures

3.1 Seven Desirable Properties

In this section we examine a number of specific properties that, as we argue, an EMQ should enjoy. The spirit of our discussion will be essentially *normative*, i.e., we will argue whether an EMQ should or should not enjoy a given property, and whether this should hold regardless of the intended application. This is different, e.g., from the spirit of (Amigó et al. 2011) (a work on the properties of evaluation measures for document filtering), which has a *descriptive* intent, i.e., describes a number of properties that such evaluation measures may or may not enjoy but does not necessarily argue that all measures should satisfy them.

The first four properties for EMQs that we discuss concern both mathematical "well-formedness" and ease of interpretation.

Property 1 ***Identity of Indiscernibles (IoI).*** *For each codeframe $\mathcal{C}$, true distribution p, and predicted distribution $\hat{p}$, it holds that $D(p, \hat{p}) = 0$ if and only if $\hat{p} = p$.* □

Property 2 ***Non-Negativity (NN).*** *For each codeframe $\mathcal{C}$, true distribution p, and predicted distribution $\hat{p}$, it holds that $D(p, \hat{p}) \geq 0$.* □

Imposing that an EMQ enjoys **IoI** and **NN** is reasonable, since altogether they indicate a score for the *perfect estimator* (defined as the estimator $\hat{p}$ such that $\hat{p} = p$) and stipulate that any other (non-perfect) estimator must obtain a score strictly higher than it; both prescriptions fit our understanding of D

[4] Note that two distributions $p(c)$ and $\hat{p}(c)$ over $\mathcal{C}$ are essentially two nonnegative-valued, length-normalized vectors of dimensionality $|\mathcal{C}|$. The literature on EMQs thus obviously intersects the literature on functions for computing the similarity of two vectors.

as a measure or error. In mathematics, a function of two probability distributions that enjoys **IoI** and **NN** (two properties that, together, are often called *Positive Definiteness*) is called a *divergence* (a.k.a. "contrast function").[5]

Property 3 ***Strict Monotonicity (MON).*** *For each codeframe $\mathcal{C}$ and true distribution p, if there are predicted distributions $\hat{p}', \hat{p}''$ and classes $c_1, c_2 \in \mathcal{C}$ such that $\hat{p}'$ and $\hat{p}''$ only differ for the fact that $\hat{p}''(c_1) < \hat{p}'(c_1) \leq p(c_1)$ and $\hat{p}''(c_2) > \hat{p}'(c_2) \geq p(c_2)$, with $|\hat{p}''(c_1) - \hat{p}'(c_1)| = |\hat{p}''(c_2) - \hat{p}'(c_2)|$, then it holds that $D(p, \hat{p}') < D(p, \hat{p}'')$.* □

If D satisfies **MON**, this means that, all other things being equal, a higher prediction error on a class c_1 (obviously matched by a higher prediction error, of opposite sign, on another class c_2) implies a higher quantification error as measured by D.

Property 4 ***Maximum (MAX).*** *There is a real value $\beta > 0$ such that, for each codeframe $\mathcal{C}$ and for each true distribution p, (i) there is a predicted distribution $\hat{p}^*$ such that $D(p, \hat{p}^*) = \beta$, and (ii) for no predicted distribution $\hat{p}$ it holds that $D(p, \hat{p}) > \beta$.* □

An estimator $\hat{p}^*$ that is the worst possible estimator of p for D (i.e., $\hat{p}^* = \arg\max_{\hat{p} \in \mathcal{P}} D(p, \hat{p})$) will be called the *perverse estimator* of p for D. If D satisfies **MAX** and $\hat{p}^*$ is the perverse estimator of p for D, then $D(p, \hat{p}^*) = \beta$. Without loss of generality, in the rest of this paper we will assume $\beta = 1$; this assumption is unproblematic since any interval $[0, \beta]$ can be rescaled to the $[0, 1]$ interval.

Altogether, these first four properties state (among other things) that the range of an EMQ that satisfies them is *independent of the problem setting* (i.e., of $\mathcal{C}$, of its cardinality $|\mathcal{C}|$, and of the true distribution p).[6] This is important, since in order to be able to *easily* judge whether a given value of D means high or low quantification error, not only we need to know what values D ranges on, but we need to know that these values are always the same. In other words, should this range depend on $\mathcal{C}$, or on its cardinality, or on the true distribution p, we would not be able to easily interpret the meaning of a given value of D.

An additional, possibly even more important reason for requiring this range to be independent of the problem setting is that, in order to test a given quantification method, the EMQ usually needs to be evaluated on a set of n test samples $\sigma_1, ..., \sigma_n$ (each characterized by its own true distribution), and a measure of central tendency (typically: the average or the median) across the n resulting EMQ values then needs to be computed (see Section 5.3 for more

[5] A divergence is often indicated by the notation $D(p||\hat{p})$); we will prefer the more neutral notation $D(p, \hat{p})$. Note also that a divergence can take as arguments any two distributions p and q defined on the same space of events, i.e., p and q need not be a true distribution and a predicted distribution. However, since we will consider divergences only as measures of fit between a true distribution and a predicted distribution, we will use the more specific notation $D(p, \hat{p})$ rather than the more general $D(p, q)$.

[6] By the "range" of an EMQ here we actually mean its *image* (i.e., the set of values that the EMQ actually takes for its admissible input values), and not just its codomain.

on this). If, for these n samples, the EMQ ranges on n different intervals, this measure of central tendency will return unreliable results, since the results obtained on the samples characterized by the wider such intervals will exert a higher influence on the resulting value.

The fifth property we discuss deals with the relative impact of underestimation and overestimation.

Property 5 ***Impartiality (IMP).*** *For any codeframe $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$, true distribution p, predicted distributions $\hat{p}'$ and $\hat{p}''$, classes $c_1, c_2 \in \mathcal{C}$, and constant $a \geq 0$ such that $\hat{p}'$ and $\hat{p}''$ only differ for the fact that $\hat{p}'(c_1) = p(c_1) + a$, $\hat{p}'(c_2) = p(c_2) - a$, $\hat{p}''(c_1) = p(c_1) - a$, $\hat{p}''(c_2) = p(c_2) + a$, it holds that $D(p, \hat{p}') = D(p, \hat{p}'')$.* □

In a nutshell, for an EMQ D that enjoys **IMP**, underestimating a true prevalence $p(c)$ by an amount a or overestimating it by the same amount a are equally serious mistakes. For instance, assume that $\mathcal{C} = \{c_1, c_2\}$, $p(c_1) = 0.10$, $p(c_2) = 0.90$, and let $\hat{p}'$ and $\hat{p}''$ be two predicted distributions such that $\hat{p}'(c_1) = 0.05$, $\hat{p}'(c_2) = 0.95$, $\hat{p}''(c_1) = 0.15$, and $\hat{p}''(c_2) = 0.85$. If an EMQ D satisfies **IMP** then $D(p, \hat{p}') = D(p, \hat{p}'')$.

We contend that **IMP** is indeed a desirable property of any EMQ, since underestimation and overestimation should be equally penalized, unless there is a specific reason for not doing so.[7] If, in a given application, we want to state that the two mistakes bring about different costs, we should be able to explicitly state these costs as parameters of the adopted measure.[8] However, in the absence of any such explicit statement, the two errors should be considered equally serious.

A further reason for insisting that an EMQ satisfies **IMP** is that the parameters of a quantifier trained via supervised learning, if optimized on a measure D that penalizes (say) the underestimation of $p(c)$ less than it penalizes its overestimation, will be such that the quantifier will systematically tend to underestimate $p(c)$. Depending on the type of parameters, this may be the result of optimization carried out either implicitly (i.e., via supervised learners that use D as the loss to minimize – see e.g., (Esuli and Sebastiani 2015)) or explicitly (i.e., via k-fold cross validation).

So far we have discussed properties that, as we claim, should be enjoyed by any EMQ. This is not the case for the next (and last) two properties since

[7] One might argue that underestimating the prevalence of a class c_1 always implies overestimating the prevalence of another class c_2. However, there are cases in which c_1 and c_2 are not equally important. For instance, if $\mathcal{C} = \{c_1, c_2\}$, with c_1 the class of patients that suffer from a certain rare disease (say, one such that $p(c_1) = .0001$) and c_2 the class of patients who do not, the class whose prevalence we really want to quantify is c_1, the prevalence of c_2 being derivative. So, what we really care about is that underestimating $p(c_1)$ and overestimating $p(c_1)$ are equally penalized. The formulation of **IMP**, which involves underestimation and overestimation in a perfectly symmetric way, is strong enough that **IMP** is not satisfied (as we will see in Section 4) by a number of important EMQs.

[8] In this case we enter the realm of *cost-sensitive quantification*, which is outside the scope of this paper; see (Forman 2008, §4&§5) and (González et al. 2017, §10) for more on the relationships between quantification and cost.

they exclude each other (i.e., an EMQ may not enjoy them both). We will claim that in some application contexts the former is desirable while in other application contexts the latter is desirable.

Property 6 ***Relativity (REL)****. For any codeframe $\mathcal{C}$, constant $a > 0$, true distributions p' and p'' that only differ for the fact that, for classes c_1 and c_2, $p'(c_1) < p''(c_1)$ and $p''(c_2) < p'(c_2)$ (with $p''(c_1) < p''(c_2)$), if a predicted distribution $\hat{p}'$ that estimates p' is such that $\hat{p}'(c_1) = p'(c_1) \pm a$ and a predicted distribution $\hat{p}''$ that estimates p'' is such that $\hat{p}''(c_1) = p''(c_1) \pm a$, and $\hat{p}'(c) = \hat{p}''(c)$ for all $c \notin \{c_1, c_2\}$, then it holds that $D(p', \hat{p}') > D(p'', \hat{p}'')$.* □

In order to understand this fairly complex formulation[9] let us see a concrete example.

Example 1 *Assume that $\mathcal{C} = \{c_1, c_2, c_3, c_4\}$, and that $p', p'', \hat{p}', \hat{p}''$ are described by the following table:*

	c_1	c_2	c_3	c_4
p'	*0.15*	*0.35*	*0.40*	*0.10*
$\hat{p}'$	*0.10*	*0.55*	*0.30*	*0.05*
p''	*0.20*	*0.30*	*0.40*	*0.10*
$\hat{p}''$	*0.15*	*0.50*	*0.30*	*0.05*

*This scenario is characterized by the fact that, of the only two classes (c_1 and c_2) that have different prevalence in p' and p'', the one with the smallest true prevalence (c_1) in both p' and p'' is underestimated by the same amount (0.05) by both $\hat{p}'$ and $\hat{p}''$. In this case D penalizes (if it satisfies **REL**) $\hat{p}'$ more than it penalizes $\hat{p}''$, since $p'(c_1) < p''(c_1)$.* □

The rationale of **REL** is that an EMQ that satisfies it, sanctions that an error of absolute magnitude a is more serious when the true class prevalence is smaller. **REL** may be a desirable property in some applications of quantification. Consider, as an example, the case in which the prevalence $p(c)$ of pathology c (say, Tubercolosis) as a cause of death in a population has to be estimated, for epidemiological purposes, from verbal descriptions of the symptoms that the deceased exhibited before dying (King and Lu 2008). In this case, **REL** should arguably be a property of the EMQ; in fact, predicting $\hat{p}'(c) = 0.0101$ when $p'(c) = 0.0001$ is a much more serious mistake than predicting $\hat{p}''(c) = 0.1100$ when $p''(c) = 0.1000$, since in the former case a very rare cause of death is overestimated by two orders of magnitude (e.g., the presence of an epidemic might mistakenly be inferred), while the same is not true in the latter case.

However, in other applications of quantification **REL** may be undesirable. To see this, consider an example in which we want to predict the prevalence $p(\text{NoShow})$ of the NoShow class among the passengers booked on a flight with actual capacity X (so that the airline can "overbook" additional

[9] The symbol $\pm$ stands for "plus or minus" while $\mp$ stands for "minus or plus"; when symbol $\pm$ evaluates to +, symbol $\mp$ evaluates to -, and vice versa.

$\hat{p}(\mathsf{NoShow}) \cdot X$ seats). In this application, relativity should arguably *not* be a property of the evaluation measure, since predicting $\hat{p}(\mathsf{NoShow}) = 0.05$ when $p(\mathsf{NoShow}) = 0.10$ or predicting $\hat{p}(\mathsf{NoShow}) = 0.15$ when $p(\mathsf{NoShow}) = 0.20$ brings about the same cost to the airline (i.e., that $0.05 \cdot X$ seats will remain empty). Applications such as this demand that the EMQ satisfies instead the following property.

Property 7 ***Absoluteness (ABS).*** *For any codeframe $\mathcal{C}$, constant $a > 0$, true distributions p' and p'' that only differ for the fact that, for classes c_1 and c_2, $p'(c_1) < p''(c_1)$ and $p''(c_2) < p'(c_2)$ (with $p''(c_1) < p''(c_2)$), if a predicted distribution $\hat{p}'$ that estimates p' is such that $\hat{p}'(c_1) = p'(c_1) \pm a$ and a predicted distribution $\hat{p}''$ that estimates p'' is such that $\hat{p}''(c_1) = p''(c_1) \pm a$, and $\hat{p}'(c) = \hat{p}''(c)$ for all $c \notin \{c_1, c_2\}$, then it holds that $D(p', \hat{p}') = D(p'', \hat{p}'')$.* □

The formulation of **ABS** only differs from the formulation of **REL** for its conclusion: while **REL** stipulates that $D(p', \hat{p}')$ must be higher than $D(p'', \hat{p}'')$, **ABS** states that the two must be equal. The rationale of **ABS** is to guarantee that an error of the same magnitude has the same impact on D regardless of the true prevalence of the class. **ABS** and **REL** are thus mutually exclusive.

Note that **ABS** and **REL** are not redundant, i.e., they do not cover the entire spectrum of possibilities (see Section 4.6 for an example EMQ that enjoys neither). For instance, an EMQ might consider an error more serious when the true class prevalence is *larger*, in which case it would satisfy neither **REL** nor **ABS**. As the two examples above show, there are applications that positively demand **REL** to hold and others that positively demand **ABS**. As a result, we will not claim that an EMQ must (or must not) enjoy **REL** or **ABS**; we simply think it is important to ascertain whether a given EMQ satisfies **REL** or **ABS** or neither, since depending on this the EMQ may or may not be adequate for the application one is tackling.

3.2 Reformulating **MON**, **IMP**, **REL**, **ABS**

The formulations of four of the properties presented above (namely, **MON**, **IMP**, **REL**, **ABS**) might seem baroque, i.e., not as tight as they could be. In this section we will try to simplify them, but for this we need to discuss a further property. In this section we will define simplified versions of them, and show that if an EMQ satisfies the **IND** property, that we are going to define next, then each of **MON**, **IMP**, **REL**, **ABS** is equivalent to its simplified counterpart. Since, as it turns out, all the measures that we discuss in this paper satisfy **IND**, this will substantially simplify the task of checking whether our measures satisfy **MON**, **IMP**, **REL**, **ABS**.

Assume a codeframe $\mathcal{C} = \{c_1, ..., c_n\}$ partitioned into $\mathcal{C}_1 = \{c_1, ..., c_k\}$ and $\mathcal{C}_2 = \{c_{k+1}, ..., c_n\}$, and a true distribution p on $\mathcal{C}$ such that $\sum_{c \in \mathcal{C}_1} p(c) = a$ for some constant $0 < a \leq 1$. We define the *projection* of p on $\mathcal{C}_1$ as the distribution $p_{\mathcal{C}_1}$ on $\mathcal{C}_1$ such that $p_{\mathcal{C}_1}(c) = \frac{p(c)}{a}$ for all $c \in \mathcal{C}_1$.

Example 2 *Assume that* $\mathcal{C} = \{c_1, c_2, c_3, c_4\}$*, that* $\mathcal{C}_1 = \{c_1, c_2, c_3\}$*, and that* p *is as in the 1st row of the following table. The* projection *of* p *on* $\mathcal{C}_1$ *is then described in the 2nd row of the same table.*

	c_1	c_2	c_3	c_4
p	*0.32*	*0.00*	*0.48*	*0.20*
$p_{\mathcal{C}_1}$	*0.40*	*0.00*	*0.60*	—

□

Essentially, the projection on $\mathcal{C}_1 \subset \mathcal{C}$ of a distribution p defined on $\mathcal{C}$ is a distribution defined on $\mathcal{C}_1$ such that the ratios between prevalences of classes that belong to $\mathcal{C}_1$ are the same in $\mathcal{C}$ and $\mathcal{C}_1$.

We are now ready to describe Property 8.

Property 8 ***Independence*** ***(IND)****. For any codeframes* $\mathcal{C} = \{c_1, ..., c_n\}$*,* $\mathcal{C}_1 = \{c_1, ..., c_k\}$ *and* $\mathcal{C}_2 = \{c_{k+1}, ..., c_n\}$*, for any true distribution* p *on* $\mathcal{C}$ *and predicted distributions* $\hat{p}'$ *and* $\hat{p}''$ *on* $\mathcal{C}$ *such that* $\hat{p}'(c) = \hat{p}''(c)$ *for all* $c \in \mathcal{C}_2$*, it holds that* $D(p, \hat{p}') \leq D(p, \hat{p}'')$ *if and only if* $D(p_{\mathcal{C}_1}, \hat{p}'_{\mathcal{C}_1}) \leq D(p_{\mathcal{C}_1}, \hat{p}''_{\mathcal{C}_1})$*.* □

If D satisfies property **IND**, this essentially means that when two predicted distributions estimate the prevalence of all classes $\{c_{k+1}, ..., c_n\}$ identically, according to D their relative merit is independent from these classes, and can thus be established by focusing only on the remaining classes $\{c_1, ..., c_k\}$.

We can now attempt to simplify the formulation of the **MON**, **IMP**, **REL**, **ABS** properties. For this discussion we will take **MON** as an example, since similar considerations also apply to the other three properties.

What we would like from a monotonicity property is to stipulate that any even small increase in quantification error must generate an increase in the value of $D(p, \hat{p})$. However, the notion of an "increase in quantification error" is non-trivial. To see this, note that characterizing an increase in *classification* error is simple, since the units of classification (the unlabelled items) are independent of each other: in a single-label context, to generate an increase in classification error one just needs to switch the predicted label of a single test items from correct to incorrect, and the other items are not affected.[10] In a quantification context, instead, increasing the difference between $p(c_i)$ and $\hat{p}(c_i)$ for some c_i does not necessarily increase quantification error, since the estimation(s) of some other class(es) in $\mathcal{C}/\{c_i\}$ is/are affected too, in many possible ways; in some cases the quantification error across the entire codeframe $\mathcal{C}$ unequivocally increases, while in some other cases it is not clear whether this happens or not, as the following example shows.

Example 3 *Assume that* $\mathcal{C} = \{c_1, c_2, c_3, c_4\}$*, and assume the following true distribution* p *and predicted distributions* $\hat{p}', \hat{p}'', \hat{p}'''$*:*

[10] This is the basis of the "Strict Monotonicity" property discussed in (Sebastiani 2015) for the evaluation of classification systems.

	c_1	c_2	c_3	c_4
p	*0.20*	*0.30*	*0.25*	*0.25*
$\hat{p}'$	*0.25*	*0.15*	*0.30*	*0.30*
$\hat{p}''$	*0.35*	*0.15*	*0.25*	*0.25*
$\hat{p}'''$	*0.35*	*0.05*	*0.30*	*0.30*

In switching from $\hat{p}'$ to $\hat{p}''$ the quantification error on c_1 increases, but the quantification error on c_3 and c_4 decreases, so that it is not clear whether we should consider the quantification error on $\mathcal{C}$ to increase or decrease. Conversely, in switching from $\hat{p}'$ to $\hat{p}'''$ the quantification errors on c_1 and *on the rest of the codeframe as a whole both increase.* □

Example 3 shows that the increase in the quantification error on a single class says nothing about how the quantification error on the entire codeframe varies. As a result, in **MON** we cannot stipulate (as we would have liked) that, in switching from one predicted distribution to another, D should increase with the increase in the estimation error on a single class c_1. The only thing we can do is to impose a monotonicity condition on how D behaves in a specific case, i.e., when the increase in the estimation error on a class c_1 is exactly matched by an estimation error (of identical magnitude but opposite sign) on another class c_2 (which is what **MON** does) while the estimation errors on all the other classes do not change.

The two predicted distributions $\hat{p}'$ and $\hat{p}''$ mentioned in **MON** are such that $\hat{p}'(c_1) + \hat{p}'(c_2) = \hat{p}''(c_1) + \hat{p}''(c_2) = a$ for some constant $0 < a \leq 1$, while both $\sum_{c \in \mathcal{C}/\{c_1,c_2\}} \hat{p}'(c)$ and $\sum_{c \in \mathcal{C}/\{c_1,c_2\}} \hat{p}''(c)$ are equal to $(1-a)$. This means that, assuming that D satisfies **IND**, we can reformulate **MON** in a way that disregards classes other than $\{c_1, c_2\}$ and considers instead the projection of p on $\{c_1, c_2\}$. In other words, if D satisfies **IND** we can reformulate **MON** in a way that tackles the problem in a binary quantification context (instead of the more general single-label quantification context). The fact that, in a binary context, $p(c_2) = (1 - p(c_1))$ for any (true or predicted) distribution p, means that **MON** can be reformulated by simply referring to just one of the two classes, i.e.,

Property 9 ***Binary Strict Monotonicity (B-MON)****. For any codeframe $\mathcal{C} = \{c_1, c_2\}$ and true distribution p, if predicted distributions $\hat{p}', \hat{p}''$ are such that $\hat{p}''(c_1) < \hat{p}'(c_1) \leq p(c_1)$, then it holds that $D(p, \hat{p}') < D(p, \hat{p}'')$.* □

As a result of what we have said in this section, **B-MON** is, for any EMQ D that satisfies **IND**, equivalent to **MON**. It is also much more compact since, among other things, it makes reference to a single class only. Considerations analogous to the ones above can be made for **IMP**, **REL**, **ABS**. We reformulate them too as below.

Property 10 ***Binary Impartiality (B-IMP)****. For any codeframe $\mathcal{C} = \{c_1, c_2\}$, true distribution p, predicted distributions $\hat{p}'$ and $\hat{p}''$, and constant $a \geq 0$ such that $\hat{p}'(c_1) = p(c_1) + a$ and $\hat{p}''(c_1) = p(c_1) - a$, it holds that $D(p, \hat{p}') = D(p, \hat{p}'')$.* □

Property 11 ***Binary Relativity (B-REL)****. For any codeframe $\mathcal{C} = \{c_1, c_2\}$, constant $a > 0$, true distributions p' and p'' such that $p'(c_1) < p''(c_1)$ and $p''(c_1) < p''(c_2)$, if a predicted distribution $\hat{p}'$ that estimates p' is such that $\hat{p}'(c_1) = p'(c_1) \pm a$ and a predicted distribution $\hat{p}''$ that estimates p'' is such that $\hat{p}''(c_1) = p''(c_1) \pm a$, then it holds that $D(p', \hat{p}') > D(p'', \hat{p}'')$.* □

Property 12 ***Binary Absoluteness (B-ABS)****. For any codeframe $\mathcal{C} = \{c_1, c_2\}$, constant $a > 0$, true distributions p' and p'' such that $p'(c_1) < p''(c_1)$ and $p''(c_1) < p''(c_2)$, if a predicted distribution $\hat{p}'$ that estimates p' is such that $\hat{p}'(c_1) = p'(c_1) \pm a$ and a predicted distribution $\hat{p}''$ that estimates p'' is such that $\hat{p}''(c_1) = p''(c_1) \pm a$, then it holds that $D(p', \hat{p}') = D(p'', \hat{p}'')$.* □

In the next sections, instead of trying to prove that an EMQ verifies Properties 3–7, we will equivalently (i) try to prove that it verifies **IND**, and if successful (ii) try to prove that it verifies Properties 9–12; the reason is, of course, the much higher simplicity and compactness of the formulations of Properties 9–12 with respect to Properties 3–7.

4 Evaluation Measures for Single-Label Quantification

In this section we turn to the functions that have been proposed and used for evaluating quantification, and discuss whether they comply or not with the properties that we have discussed in Section 3. In many cases these functions were originally proposed for evaluating the binary case; since the extension to SLQ is usually straightforward, for each EMQ we indicate its original proponent or user (on this see also Table 2) and disregard whether it was originally used just for BQ or for the full-blown SLQ.

We will discuss 9 measures proposed as EMQs in the literature, and for each of them we will be interested in whether they satisfy or not Properties 1 to 8. Giving $9 \times 8 = 72$ proofs in detail would make the paper excessively long and boring: as a result, only some of these proofs will be given in detail, while for others we will only give hints at how they can be easily obtained via the same lines of reasoning used in other cases. In several cases, given a measure D and a property π, one can simply show that D does *not* enjoy π via a counterexample. Since the same scenario can serve as a counterexample for showing that π is not enjoyed by several measures, we formulate each such scenario in the form of a table that shows which measures the scenario rules out. In the appendix we include a table each for properties **MAX** (Appendix B.1), **IMP** (Appendix B.2), **REL** (Appendix B.3), **ABS** (Appendix B.4); in this section, when discussing the property in the context of a specific measure that does not enjoy it, we will simply refer the reader to the appropriate table.

A 2D plot (for the case of binary quantification) of the 9 measures we will discuss is displayed in Figure 1; Figure 2 displays the same plots in 3D. These plots allow to graphically appreciate if a measure enjoys a certain property or not. For instance, looking at the 2D plots, a measure that enjoys both **IoI** and **NN** (i.e., a divergence) is such that the $y = x$ diagonal is the locus of

the darkest points; a measure that enjoys **MON** is such that, when moving away in a vertical direction (i.e., up or down) from the $y = x$ diagonal, points get lighter; a measure that enjoys **IMP** is such that moving away in a vertical direction from the $y = x$ diagonal, moving up or down by the same amount returns points of the same colour; a measure that enjoys **ABS** is such that moving away in a vertical direction from the $y = x$ diagonal in a given sense (e.g., down), the difference in colour does not depend from which point of the diagonal we are moving away from; etc.

4.1 Absolute Error

The simplest EMQ is *Absolute Error* (AE), which corresponds to the average (across the classes in $\mathcal{C}$) absolute difference between the predicted class prevalence and the true class prevalence; i.e.,

$$\mathrm{AE}(p, \hat{p}) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} |\hat{p}(c) - p(c)| \tag{1}$$

It is easy to prove that AE enjoys **IoI**, **NN**, **MON**, **IMP**, **ABS**, **IND**. While some of these proofs are trivial, we report them in detail (in Appendix A) in order to show how the same arguments can be used to prove the same for many of the EMQs to be discussed later in this section.

Instead, as shown in Appendix B.1, AE does not enjoy **MAX**, because its range depends on the true distribution p. More specifically, AE ranges between 0 (best) and

$$z_{\mathrm{AE}} = \frac{2(1 - \min_{c \in \mathcal{C}} p(c))}{|\mathcal{C}|} \tag{2}$$

(worst), i.e., its range depends also on the cardinality of $\mathcal{C}$. In fact, it is easy to verify that, given a true distribution p on $\mathcal{C}$, the perverse estimator of p is the one such that (a) $\hat{p}(c^*) = 1$ for class $c^* = \arg\min_{c \in \mathcal{C}} p(c)$, and (b) $\hat{p}(c) = 0$ for all $c \in \mathcal{C}/\{c^*\}$. In this case, the *total* error derives (i) from overestimating $p(c^*)$, which brings about an error of $(1-p(c^*))$, and (ii) from underestimating $p(c)$ for all $c \in \mathcal{C}/\{c^*\}$, which collectively brings about an additional error of $(1 - p(c^*))$. AE is obtained by dividing this $2(1 - p(c^*))$ quantity by $|\mathcal{C}|$.

Concerning **REL**, just note that since AE satisfies **ABS**, it cannot (as observed in Section 3) satisfy **REL**. (That AE does not enjoy **REL** is also shown via a counterexample in Appendix B.3.)

The properties that AE enjoys (and those it does not enjoy) are conveniently summarized in Table 1, along with the same for all the measures discussed in the rest of this paper.

In the literature, AE also goes by the name of *Variational Distance* (Csiszár and Shields 2004, §4),(Lin 1991; Zhang and Zhou 2010), or *Percentage Discrepancy* (Esuli and Sebastiani 2010; Baccianella et al. 2013). Also, if viewed

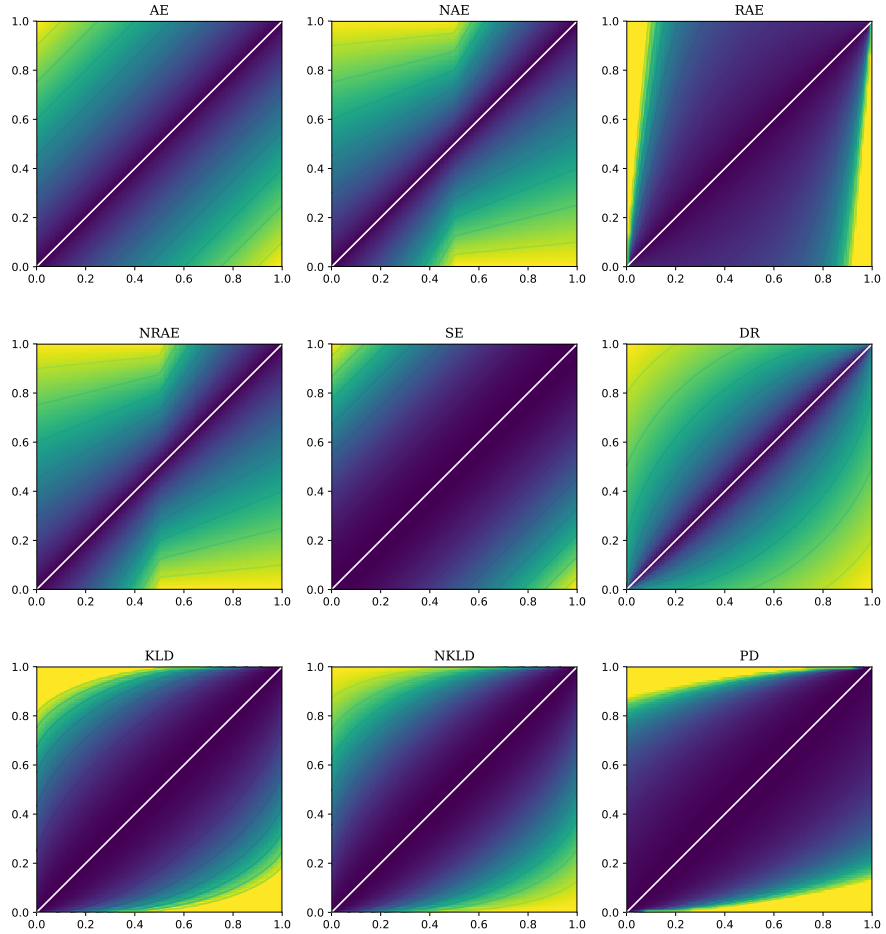

Fig. 1 2D plots (for a binary quantification task) for the nine EMQs of Tables 1 and 2; $p(c_1)$ and $p(c_2)$ are represented as x and $(1-x)$, respectively, while $\hat{p}(c_1)$ and $\hat{p}(c_2)$ are represented as y and $(1-y)$. Darker areas represent values closer to 0 (i.e., smaller error) while lighter areas represent values more distant from 0 (i.e., higher error).

as a generic function of dissimilarity between vectors (and not just probability distributions), AE is nothing else than the well-known "city-block distance" normalized by the number of classes. Some recent papers (Beijbom et al. 2015; González et al. 2017) that tackle quantification in the context of ecological modelling discuss or use, as an EMQ, *Bray-Curtis dissimilarity* (BCD), a measure popular in ecology for measuring the dissimilarity of two samples. However, when used to measure the dissimilarity of two probability distributions, BCD defaults to AE; as a result we will not analyse BCD any further.

Note that AE often goes by the name of *Mean* Absolute Error; for simplicity, for this and the other measures we discuss in the rest of this paper we will omit the qualification "Mean", since every measure mediates across the class-specific values in its own way.

As an EMQ, AE was used for the first time by (Saerens et al. 2002), and in many other papers ever since. For AE and for all the other EMQs discussed in this paper, Table 2 lists the papers where the measure has been proposed and those which have subsequently used it for evaluation purposes.

4.2 Normalized Absolute Error

Following what we have said in Section 4.1, a normalized version of AE that always ranges between 0 (best) and 1 (worst) can be obtained as

$$\mathrm{NAE}(p, \hat{p}) = \frac{\mathrm{AE}(p, \hat{p})}{z_{\mathrm{AE}}} = \frac{\sum_{c \in \mathcal{C}} |\hat{p}(c) - p(c)|}{2(1 - \min_{c \in \mathcal{C}} p(c))} \tag{3}$$

where z_{AE} is as in Equation 2. It is easy to verify that NAE enjoys **IoI**, **NN**, **MON**, **IMP**, **IND**. NAE also enjoys (by construction) **MAX**.

Given that NAE is just a normalized version of AE, and given that AE enjoys **ABS**, one might expect that NAE enjoys **ABS** too. Surprisingly enough, this is not the case, as shown in the counterexample of Appendix B.4. The reason for this is that, for the two distributions p' and p'' (and their respective predicted distributions $\hat{p}'$ and $\hat{p}''$) mentioned in the formulation of Property 7 (**ABS**), and exemplified in the counterexample of Appendix B.4, the numerator of Equation 3 is the same but the denominator (i.e., the normalizing constant) is different, which means that the value of NAE is also different. NAE does not enjoy **REL** either, as also shown in Appendix B.3).

NAE was discussed for the first time by Esuli and Sebastiani (2014). With a similar intent, in a binary quantification context Barranquero et al. (2015) proposed *Normalized Absolute Score* (NAS). NAS is an accuracy (and not an error) measure; when viewed as an error measure, it is defined as

$$\mathrm{NAS}(p, \hat{p}) = \frac{|p(c) - \hat{p}(c)|}{\max\{p(c), (1 - p(c))\}} \tag{4}$$

where c is any class in $\mathcal{C} = \{c_1, c_2\}$. We will not discuss NAS in detail since (a) it is only defined for the binary case, and (b) it is easy to show that in this case it coincides with NAE.

4.3 Relative Absolute Error

Relative Absolute Error (RAE) relativises the value $|\hat{p}(c) - p(c)|$ in Equation 1 to the true class prevalence, i.e.,

$$\mathrm{RAE}(p, \hat{p}) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{|\hat{p}(c) - p(c)|}{p(c)} \tag{5}$$

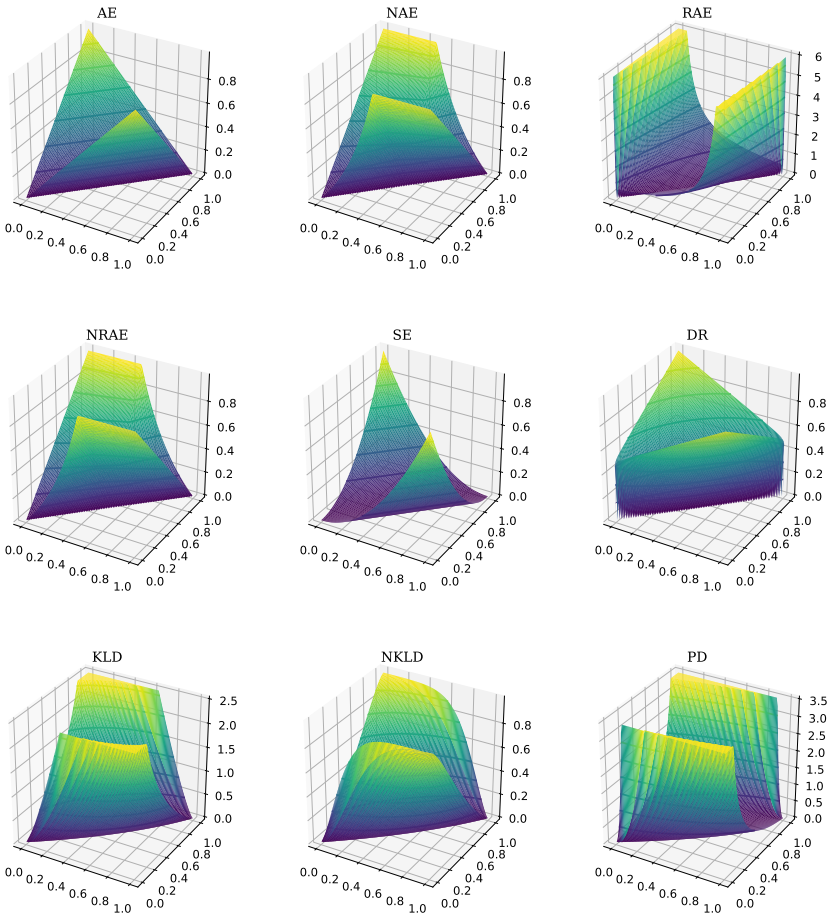

Fig. 2 3D plots (for a binary quantification task) for the nine EMQs of Tables 1 and 2; $p(c_1)$ and $p(c_2)$ are represented as x and $(1-x)$, respectively, while $\hat{p}(c_1)$ and $\hat{p}(c_2)$ are represented as y and $(1-y)$; error is represented as z (higher values of z represent higher error).

RAE may be undefined in some cases, due to the presence of zero denominators. To solve this problem, in computing RAE we can smooth both $p(c)$ and $\hat{p}(c)$ via additive smoothing, i.e., we take

$$p_s(c) = \frac{\epsilon + p(c)}{\epsilon|\mathcal{C}| + \sum_{c \in \mathcal{C}} p(c)} \tag{6}$$

where $p_s(c)$ denotes the smoothed version of $p(c)$ and the denominator is just a normalizing factor (same for the $\hat{p}_s(c)$'s); the quantity $\epsilon = \frac{1}{2|\sigma|}$ is often used (and will always be used in the rest of this paper) as a smoothing factor. The

smoothed versions of $p(c)$ and $\hat{p}(c)$ are then used in place of their original non-smoothed versions in Equation 5; as a result, RAE is always defined.

Using arguments analogous to the ones used for AE in Appendix A, it is immediate to show that RAE enjoys **IoI**, **NN**, **MON**, **IMP**, **IND**. It also enjoys **REL** by construction, which means that it does not enjoy **ABS**. Analogously to AE, the fact RAE does not enjoy **MAX**, as shown via the counterexample in Appendix B.1.

It is easy to show that RAE ranges between 0 (best) and

$$z_{\text{RAE}} = \frac{|\mathcal{C}| - 1 + \frac{1 - \min_{c \in \mathcal{C}} p(c)}{\min_{c \in \mathcal{C}} p(c)}}{|\mathcal{C}|} \tag{7}$$

(worst), i.e., its range depends also on the cardinality of $\mathcal{C}$. In fact, similarly to the case of AE, it is easy to verify that, given a true distribution p on $\mathcal{C}$, the perverse estimator of p is obtained when (a) $\hat{p}(c) = 1$ for the class $c^* = \arg\min_{c \in \mathcal{C}} p(c)$, and (b) $\hat{p}(c) = 0$ for all $c \in \mathcal{C}/\{c^*\}$. In this case, the *total* relative absolute error derives (i) from overestimating $p(c^*)$, which brings about an error of $\frac{1-p(c^*)}{p(c^*)}$, and (ii) from underestimating $p(c)$ for all $c \in \mathcal{C}/\{c^*\}$, which brings about an additional error of 1 for each class in $\mathcal{C}/\{c^*\}$. The value of RAE is then obtained by dividing the resulting $(|\mathcal{C}| - 1 + \frac{1 - p(c^*)}{p(c^*)})$ by $|\mathcal{C}|$.

As an EMQ, RAE was used for the first time by González-Castro et al. (2010), and by several other papers after it.

4.4 Normalized Relative Absolute Error

Following what we have said in Section 4.3, a normalized version of RAE that always ranges between 0 (best) and 1 (worst) can thus be obtained as

$$\text{NRAE}(p, \hat{p}) = \frac{\text{RAE}(p, \hat{p})}{z_{\text{RAE}}} = \frac{\sum_{c \in \mathcal{C}} \frac{|\hat{p}(c) - p(c)|}{p(c)}}{|\mathcal{C}| - 1 + \frac{1 - \min_{c \in \mathcal{C}} p(c)}{\min_{c \in \mathcal{C}} p(c)}} \tag{8}$$

where z_{RAE} is as in Equation 7. Since the various denominators of Equation 8 may be undefined, the smoothed values of Equation 6 must be used in Equation 8 too.

It is straightforward to verify that NRAE, which was first proposed by Esuli and Sebastiani (2014), enjoys **IoI**, **NN**, **MON**, **IMP**, **IND**, and also enjoys (by construction) **MAX**.

Somehow similarly to what we said in Section 4.2 about NAE and **ABS**, given that NRAE is just a normalized version of RAE, and given that RAE enjoys **REL**, one might expect that NRAE enjoys **REL** too. Again, this is

not the case, as shown in the counterexample of Appendix B.3. The reason for this is that, for the two distributions p' and p'' (and their respective predicted distributions $\hat{p}'$ and $\hat{p}''$) mentioned in the formulation of Property 6 (**REL**), and exemplified in the counterexample of Appendix B.3, while RAE (the numerator of Equation 8) does enjoy **REL**, the normalizing constant (the denominator of Equation 8) invalidates it, since it is different for p' and p''. NAE does not enjoy **ABS** either, as also shown in Appendix B.4.

4.5 Squared Error

Another measure that has been used in the quantification literature is *Squared Error* (SE), defined as

$$\mathrm{SE}(p,\hat{p}) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}(p(c)-\hat{p}(c))^2 \tag{9}$$

When viewed as a generic function of dissimilarity between vectors (and not just probability distributions), SE is the well-known L^2-distance. As an EMQ, SE was used for the first time by Bella et al. (2010).

The mathematical form of SE is very similar to that of AE, and it can be trivially shown that SE enjoys all the properties that AE enjoys and does not enjoy all the properties that AE does not enjoy. In particular, SE does not enjoy **MAX** since SE ranges between 0 (best) and

$$z_{\mathrm{SE}} = \frac{(1-p(c^*))^2 + \sum_{c\in\mathcal{C}/\{c^*\}} p(c)^2}{|\mathcal{C}|} \tag{10}$$

(worst), where $c^* = \arg\min_{c\in\mathcal{C}} p(c)$; i.e., the range of SE depends on p and $|\mathcal{C}|$. In fact, similarly to the case of AE, it is easy to verify that the perverse estimator of a true distribution p is the one such (a) $\hat{p}(c^*) = 1$ and (b) $\hat{p}(c) = 0$ for all $c \in \mathcal{C}/\{c^*\}$. In this case, the squared error derives (i) from overestimating $p(c^*)$, which brings about an error of $\frac{(1-p(c^*))^2}{|\mathcal{C}|}$, and (ii) from underestimating $p(c)$ for all $c \in \mathcal{C}/\{c^*\}$, which brings about an additional error of $\frac{p(c)^2}{|\mathcal{C}|}$ for each class in $\mathcal{C}/\{c^*\}$. We could thus define a normalized version of SE as

$$\mathrm{NSE}(p,\hat{p}) = \frac{\mathrm{SE}(p,\hat{p})}{z_{\mathrm{SE}}} = \frac{\sum_{c\in\mathcal{C}}(p(c)-\hat{p}(c))^2}{(1-p(c^*))^2 + \sum_{c\in\mathcal{C}/\{c^*\}} p(c)^2} \tag{11}$$

which would, quite obviously, enjoy and not enjoy exactly the same properties that NAE enjoys and does not enjoy.

SE is structurally similar to AE but (as can also be appreciated from Figure 1) is less sensitive than it, i.e., it is always the case that $\mathrm{SE}(p,\hat{p}) \leq \mathrm{AE}(p,\hat{p})$ (since it is always the case that $(p(c)-\hat{p}(c))^2 \leq |p(c)-\hat{p}(c)|$).

In the binary quantification literature, other proxies of SE have been used; one example is *Normalized Squared Score* (Barranquero et al. 2015), defined as

$\text{NSS}(p(c), \hat{p}(c)) \equiv 1 - (\frac{p(c)-\hat{p}(c)}{\max\{p(c),(1-p(c))\}})^2$, where c is any class in $\mathcal{C} = \{c_1, c_2\}$. Similarly to the NAS measure discussed at the end of Section 4.1, NSS is an accuracy (and not an error) measure; when viewed as an error measure, it would be defined as

$$\text{NSS}(p, \hat{p}) = (\frac{p(c) - \hat{p}(c)}{\max\{p(c), (1 - p(c))\}})^2 \tag{12}$$

where c is any class in $\mathcal{C} = \{c_1, c_2\}$. We will not discuss NSS in detail since (a) it is only defined for the binary case, and (b) it is easy to show that in this case it coincides with NSE.

4.6 Discordance Ratio

Levin and Roitman (2017) introduce an EMQ that they call *Concordance Ratio* (CR). CR is a measure of accuracy, and not a measure of error; for better consistency with the rest of this paper, instead of CR we consider what might be called *Discordance Ratio*, i.e., its complement DR = (1−CR), defined as

$$\begin{aligned}
\text{DR}(p, \hat{p}) &= 1 - \text{CR} \\
&= 1 - \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{\min\{(p(c), \hat{p}(c)\}}{\max\{(p(c), \hat{p}(c)\}} \\
&= \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{\max\{(p(c), \hat{p}(c)\} - \min\{(p(c), \hat{p}(c)\}}{\max\{(p(c), \hat{p}(c)\}} \\
&= \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{|(p(c) - \hat{p}(c)|}{\max\{(p(c), \hat{p}(c)\}}
\end{aligned} \tag{13}$$

DR is undefined when, for a given class c, both $p(c)$ and $\hat{p}(c)$ are zero; the smoothed values of Equation 6 must thus be used within Equation 13 in order to avoid this problem.

It is easy to verify, along the lines sketched in Appendix A, that DR enjoys **IoI**, **NN**, **MON**, **IND**. DR also enjoys **REL**; this can be seen by the fact that, for the same amount a of misprediction, $\sum_{c \in \mathcal{C}} \frac{\min\{(p(c), \hat{p}(c)\}}{\max\{(p(c), \hat{p}(c)\}}$ is smaller (hence $\text{DR}(p, \hat{p})$ is larger) when the true prevalence of the class c_1 mentioned in the formulation of Property 6 (**REL**) is smaller. Instead, DR enjoys neither **MAX**, nor **IMP**, nor **ABS**, as shown in Appendixes B.1, B.2 and B.4, respectively.

4.7 Kullback-Leibler Divergence

An EMQ that has become somehow standard in the evaluation of single-label (and, *a fortiori*, binary) quantification, is *Kullback-Leibler Divergence* (KLD –

also called *Information Divergence*, or *Relative Entropy*) (Csiszár and Shields 2004) and defined as[11]

$$\mathrm{KLD}(p, \hat{p}) = \sum_{c \in \mathcal{C}} p(c) \log \frac{p(c)}{\hat{p}(c)} \tag{14}$$

As an EMQ, KLD was used for the first time (under the name *Normalized Cross-Entropy*) by Forman (2005). It should also be noted that KLD has been adopted as the official evaluation measure of the only quantification-related shared task that has been organized so far, Subtask D "Tweet Quantification on a 2-point Scale" of SemEval-2016 and Semeval-2017 "Task 4: Sentiment Analysis in Twitter" (Nakov et al. 2016, 2017).

KLD may be undefined in some cases. While the case in which $p(c) = 0$ is not problematic (since continuity arguments indicate that $0 \log \frac{0}{a}$ should be taken to be 0 for any $a \geq 0$), the case in which $\hat{p}(c) = 0$ and $p(c) > 0$ is indeed problematic, since $a \log \frac{a}{0}$ is undefined for $a > 0$. To solve this problem, we smooth values in the same way as already described in Section 4.3; as a result, KLD is always defined.

The fact that KLD enjoys **IoI** and **NN** (i.e., the fact that KLD is indeed a divergence) is not self-evident (since $p(c) \log \frac{p(c)}{\hat{p}(c)}$ is negative whenever $p(c) < \hat{p}(c)$), and is known as *Gibbs' inequality*. A formal proof of it can be found on several information theory textbooks (see e.g., (MacKay 2003, p. 44)).

Indeed, KLD is a well-known member of the class of *f-divergences* (Ali and Silvey 1966) (Csiszár and Shields 2004, §4), a class of functions that measure the difference between two probability distributions, and that all enjoy **IoI** and **NN**.

The fact that KLD enjoys **MON** is also not self-evident, essentially for the same reasons for which it is not self-evident that it enjoys **IoI** and **NN**. The proof that KLD enjoys **MON** is given in Appendix C, where we use the fact that KLD enjoys **IND** (something which can be easily shown via the arguments used in Appendix A) and thus limit ourselves to proving that it enjoys **B-MON**.

The fact that KLD enjoys neither **MAX**, nor **IMP**, nor **REL**, nor **ABS** is shown in Appendixes B.1, B.2, B.3, B.4, respectively. Concerning **MAX** we note that, in theory, the upper bound of KLD is not finite, since Equation 14 has predicted probabilities, and not true probabilities, at the denominator. That is, by making a predicted probability $\hat{p}(c)$ infinitely small we can make KLD infinitely large. However, since we use smoothed values, the fact that both p and $\hat{p}$ are lower-bounded by ϵ, and not by 0, has the consequence that KLD has a finite upper bound. The perverse estimator for KLD is the one such (a) $\hat{p}(c^*) = 1$ and (b) $\hat{p}(c) = 0$ for all $c \in \mathcal{C}/\{c^*\}$. The value of this estimator is

$$z_{\mathrm{KLD}}(p, \hat{p}) = p_s(c^*) \log \frac{p_s(c^*)}{1 - (|\mathcal{C}| - 1) \cdot \epsilon} + \sum_{c \in \mathcal{C}/\{c^*\}} p_s(c) \log \frac{p_s(c)}{\epsilon} \tag{15}$$

[11] In Equation 14 and in the rest of this paper the log operator denotes the natural logarithm.

which shows that the range of KLD depends on p, the cardinality of $\mathcal{C}$, and even on the value of ϵ. This is a further proof that KLD does not enjoy **MAX**.

4.8 Normalized Kullback-Leibler Divergence

Given what we have said in Section 4.7, one might define a normalized version of KLD (i.e., one that also enjoys **MAX**) as $\frac{\text{KLD}(p,\hat{p})}{z_{\text{KLD}}(p,\hat{p})}$, where $z_{\text{KLD}}(p,\hat{p})$ is as in Equation 15. Esuli and Sebastiani (2014) follow instead a different route, and define a normalized version of KLD by applying to it a logistic function,[12] i.e.,[13]

$$\text{NKLD}(p,\hat{p}) = 2\frac{e^{\text{KLD}(p,\hat{p})}}{e^{\text{KLD}(p,\hat{p})}+1} - 1 \tag{16}$$

Like other previously discussed measures, also NKLD may be undefined in some cases; therefore, also in computing NKLD we need to use the smoothed values of Equation 6 in place of the original $p(c)$'s and $\hat{p}(c)$'s.

NKLD enjoys some of our properties of interest for the simple reason that KLD enjoys them; it is easy to verify that this is the case of **IoI** and **NN**. NKLD also enjoys **MON** and **IND**; this descends from the fact that $\text{NKLD}(d,d') < \text{NKLD}(d,d'')$ if and only if $\text{KLD}(d,d') < \text{KLD}(d,d'')$ (this derives from the fact that the logistic function is a monotonic transformation) and from the fact that KLD enjoys **MON** and **IND**, respectively. Concerning **MAX**, NKLD enjoys it by construction, because when a predicted prevalence $\hat{p}(c)$ tends to 0 KLD tends to $+\infty$, and NKLD thus tends to 1.[14]

The fact that KLD enjoys neither **IMP**, nor **REL**, nor **ABS**, is shown in Appendixes B.2, B.3, B.4, respectively.

4.9 Pearson Divergence

The last EMQ we discuss is the *Pearson Divergence* (PD – see (du Plessis and Sugiyama 2012)), also called the χ^2 *Divergence* (Liese and Vajda 2006), and defined as

$$\text{PD}(p,\hat{p}) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\frac{(p(c)-\hat{p}(c))^2}{\hat{p}(c)} \tag{17}$$

[12] Since the standard logistic function $\frac{e^x}{e^x+1}$ ranges (for the domain $[0,+\infty)$ we are interested in) on $[\frac{1}{2},1]$, a multiplication by 2 is applied in order for it to range on [1,2], and 1 is subtracted in order for it to range on [0,1], as desired.

[13] Esuli and Sebastiani (2014) mistakenly defined $\text{NKLD}(p,\hat{p})$ as $\frac{e^{\text{KLD}(p,\hat{p})}-1}{e^{\text{KLD}(p,\hat{p})}}$; this was later corrected into the formulation of Equation 16 (which is equivalent to $\frac{e^{\text{KLD}(p,\hat{p})}-1}{e^{\text{KLD}(p,\hat{p})}+1}$) by Gao and Sebastiani (2016).

[14] This is true only at a first approximation, though. In more precise terms, the maximum value that NKLD can have is strictly smaller than 1 because the maximum value that KLD can have is finite (see Equation 15) and, as discussed at the end of Section 4.7, dependent on p, on the cardinality of $\mathcal{C}$, and even on the value of ϵ; as a result, the maximum value that NKLD can have is also dependent on these three variables (although it is always very close to 1 – see the example in Appendix B.1).

Table 1 Properties of the EMQs discussed in this paper.

	IoI	**NN**	**MAX**	**MON**	**IMP**	**REL**	**ABS**	**IND**
AE	Yes	Yes	No	Yes	Yes	No	Yes	Yes
NAE	Yes	Yes	Yes	Yes	Yes	No	No	Yes
RAE	Yes	Yes	No	Yes	Yes	Yes	No	Yes
NRAE	Yes	Yes	Yes	Yes	Yes	No	No	Yes
SE	Yes	Yes	No	Yes	Yes	No	Yes	Yes
DR	Yes	Yes	No	Yes	No	Yes	No	Yes
KLD	Yes	Yes	No	Yes	No	No	No	Yes
NKLD	Yes	Yes	Yes	Yes	No	No	No	Yes
PD	Yes	Yes	No	Yes	No	No	No	Yes

As an EMQ, PD has been first used by Ceron et al. (2016). PD is undefined when, for a given class c, $\hat{p}(c)$ is zero; the smoothed values of Equation 6 must thus be used within Equation 17 in order to avoid this problem.

The arguments already used for AE in Appendix A can be easily used to show that PD enjoys **IoI**, **NN**, and **IND**. That PD enjoys **MON** is instead not self-evident; the proof that it indeed does is reported in Appendix C.

That PD enjoys neither **MAX**, nor **IMP**, nor **REL**, nor **ABS**, is shown in Appendixes B.1, B.2, B.3, B.4, respectively. The fact that PD does not enjoy **MAX** can also be shown with arguments used for showing the same for KLD; that is, when a predicted probability $\hat{p}(c)$ is very small, PD becomes very large. Thanks to the fact that we use smoothed values, though, $\hat{p}$ is lower-bounded by ϵ, and PD has thus a finite upper bound. Like for other EMQs we have already discussed, the perverse estimator for PD is the one that attributes 1 to the probability of class $c^* = \arg\min_{c \in \mathcal{C}} p(c)$ and 0 to the other classes, and its value is thus

$$z_{\text{PD}}(p, \hat{p}) = \frac{1}{|\mathcal{C}|}\left(\frac{1 - (|\mathcal{C}| - 1) \cdot \epsilon - p_s(c^*)}{1 - (|\mathcal{C}| - 1) \cdot \epsilon} + \sum_{c \in \mathcal{C}/\{c^*\}} \frac{(p(c) - \epsilon)^2}{\epsilon}\right) \tag{18}$$

which shows that the range of PD depends on p, the cardinality of $\mathcal{C}$, and the value of ϵ. This suffices to show that PD does not enjoy **MAX**.

5 Discussion

The properties that the EMQs of Section 4 enjoy and do not enjoy are conveniently summarized in Table 1. Table 2 lists instead the papers where the various EMQs have been proposed and the papers where they have subsequently been used for evaluation purposes.

5.1 Are all our Properties Equally Important?

An examination of Table 1 allows us to make a number of general considerations. The first one is that some of our properties (namely: **IoI**, **NN**, **MON**,

IND) are unproblematic, since all the EMQs proposed so far satisfy them, while other properties (namely: **MAX**, **IMP**, **REL**, **ABS**) are failed by several EMQs, including ones (e.g., AE, KLD) that are almost standard in the quantification literature. The second, related observation is that, if we agree on the fact that the eight properties we have discussed are desirable, a number of EMQs that have been proposed in the quantification literature emerge as severely inadequate, since they fail several among these properties; this is true even if we discount the fact that, as we have already observed, **REL** and **ABS** are mutually exclusive. The case of KLD (which fails on counts of **MAX**, **IMP**, **REL**, **ABS**) is of special significance, since KLD has almost become a standard in the evaluation of single-label (and binary) quantification (from Table 2 KLD emerges as the 2nd most frequently used EMQ, after AE).

However, an even more compelling fact that emerges from Table 1 is that no EMQ among those proposed so far satisfies (even discounting the mutual exclusivity of **REL** and **ABS**) all the proposed properties. This suggests that more research is needed in order to identify, or synthesize, an EMQ more satisfactory than all the existing ones.

At the same time, in the absence of a truly satisfactory EMQ, we think that it is important to analyse whether all of our properties are equally important, or if some of them is less important than others and can thus be "sacrificed". Judging from Table 1, the key stumbling block seems to be the **MAX** property, since all the EMQs that satisfy **MAX** (namely: NAE, NRAE, NKLD) satisfy neither **REL** nor **ABS**. This is undesirable since, as argued at the end of Section 3.1, some applications of quantification do require **REL**, while some other applications do require **ABS** (and we can think of no application that requires neither). Among the EMQs that satisfy **ABS** (and not **REL**), AE and SE satisfy all other properties but **MAX**, while among the ones that satisfy **REL** (and not **ABS**), also RAE satisfies all other properties but **MAX**.

In other words, if we stick to available EMQs, if we want **ABS** or **REL** we need to renounce to **MAX**, while if we want **MAX** we need to renounce to both **ABS** and **REL**. How relatively desirable are these three properties? We recall from Section 3.1 that

1. the argument in favour of **REL** is that it reflects the needs of applications in which an estimation error of a given absolute magnitude should be considered more serious if it affects a rarer class;
2. the argument in favour of **ABS** is that it reflects the needs of applications in which an estimation error of a given absolute magnitude should be considered to have the same impact independently from the true prevalence of the affected class;
3. the main (although not the only) argument in favour of **MAX** is that, if an EMQ does not satisfy it, the n samples on which we may want to compare our quantification algorithms will each have a different weight on the final result.

The relative importance of these three arguments is probably a matter of opinion. However, it is our impression that Arguments 1 and 2 are more compelling

than Argument 3, since 1 and 2 are really about how an evaluation measure reflects the needs of the application for which one performs a given task (quantification, in our case); if the corresponding properties are not satisfied, one may argue that the quantification accuracy (or error) being measured is only loosely related to what the user really wants.

Argument 3, while important, "only" implies that, if **MAX** is not satisfied, (1) results obtained on codeframes of different cardinality will not be comparable, and (2) results obtained on samples characterized by different true distributions will not be comparable[15]; while undesirable, this does not affect the experimental comparison among different quantification systems, since each of them is affected by these disparities in the same way.[16]

So, if we accept the idea of "sacrificing" **MAX** in order to retain **REL** or **ABS**, Table 1 indicates that our measures of choice should be

- AE (or SE, which is structurally similar), for those applications in which an estimation error of a given absolute magnitude should be considered more serious when the true prevalence of the affected class is lower; and
- RAE, for those applications in which an estimation error of a given absolute magnitude has the same impact independently from the true prevalence of the affected class.

5.2 Properties that Escape Formalization

While all the above discussion on the properties of EMQs has been unashamedly formal, we should also remember that choosing an evaluation measure instead of another should also be guided by practical considerations, i.e., by properties of the measure that are not necessarily amenable to formalization. One such property is understandability, i.e., how simple and intuitive is the mathematical form of an evaluation measure. While such simplicity might not be a primary concern for the researcher, or the mathematician, it might be for the practitioner. For instance, a company that wants to sell a text analytics product to a customer might need to run experiments on the customer's own data and explain the results to the customer; since customers might not be mathematically savvy, the fact that the measure chosen is easily understandable to people with a minimal mathematical background is important. On this account, measures such as AE and RAE certainly win over other measures such as KLD and NKLD, which the average customer would find hardly intelligible.[17]

[15] It has to be remarked that, in some cases, differences of the latter type may be moderate, especially when $|C|$ is high. For instance, when $|C| = 2$ the value of z_{AE} ranges on [0.5,1.0], but when $|C| = 10$ it ranges on [0.18,0.20].

[16] A similar situation occurs when evaluating multi-label classification via "microaveraged F_1", a measure in which the classes with higher prevalence weigh more on the final result.

[17] It is this author's experience that even measures such as F_1 can be considered by customers "esoteric".

Another property that is difficult to formalize is *robustness to outliers*. Many EMQs often take the form of an average $D(p,\hat{p}) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} f(p(c),\hat{p}(c))$ across the classes in the codeframe. If $D(p,\hat{p})$ is not "robust to outliers", it means that an extreme value $f(p(c'),\hat{p}(c'))$ that may occur for some $c' \in \mathcal{C}$ dominates on all the other values $f(p(c),\hat{p}(c))$ for $c \in \mathcal{C}/\{c'\}$, giving rise to a high value of $D(p,\hat{p})$ that is essentially due to c' only. As the name implies, "robustness to outliers" is usually considered a desirable property; however, in some contexts it might also viewed as undesirable (e.g., we might want to avoid quantification methods that generate blatant mistakes, so we might want a measure that penalizes the presence of even one of them). Aside from the fact that its desirability is questionable, it should also be mentioned that "robustness to outliers" comes in degrees. E.g., absolute error is more robust to outliers than squared error, but squared error is more robust to outliers than "cubic error", etc.; and all of them are vastly more robust to outliers than KLD and NKLD. Which among these enforces the "right" level of robustness to outliers? This shows that robustness to outliers, independently from its desirability, cannot be framed as a binary property (i.e., one that a measure either enjoys or not), and thus escapes the type of analysis that we have carried out in this paper.

Another property which is difficult to formalize has to do with the set of values which an EMQ ranges on when evaluating *realistic* quantification systems (i.e., systems that exhibit a quantification accuracy equal or superior to, say, that of a trivial "classify and count" approach using SVMs). For these systems, the actual values that an EMQ takes should occupy a fairly small subinterval of its entire range. The question is: how small? One particularly problematic EMQ, from this respect, is KLD. While its range is $[0, z_{\text{KLD}}]$, where z_{KLD} is as in Equation 15, realistic quantification systems generate *very* small KLD values, so small that they are sometimes difficult to make sense of. One result is that two genuine quantifiers that are being compared experimentally may easily obtain results several *orders of magnitude* away. Such differences in performance are difficult to grasp.[18] We should add that, if one wants to average KLD results across a set of samples (on this see also Section 5.3), the average is completely dominated by the value with the highest order of magnitude, and the others have little or no impact. Unfortunately, switching from KLD to NKLD does not help much in this respect since, for realistic quantification systems, $\text{NKLD}(p,\hat{p}) \approx \frac{1}{2}\,\text{KLD}(p,\hat{p})$. The reason is that NKLD is obtained by applying a sigmoidal function (namely, the logistic function) to KLD, and the tangent to this sigmoid for $x = 0$ is $y = \frac{1}{2}x$; since the values of KLD for

[18] As an example, assume a (very realistic) scenario in which $|\sigma| = 1000$, $\mathcal{C} = \{c_1, c_2\}$, $p(c_1) = 0.01$, and in which three different quantifiers $\hat{p}'$, $\hat{p}''$, $\hat{p}'''$ are such that $\hat{p}'(c_1) = 0.0101$, $\hat{p}''(c_1) = 0.0110$, $\hat{p}'''(c_1) = 0.0200$. In this scenario KLD ranges on $[0, 7.46]$, $\text{KLD}(p,\hat{p}') = 4.78\text{e-}07$, $\text{KLD}(p,\hat{p}'') = 4.53\text{e-}05$, $\text{KLD}(p,\hat{p}''') = 3.02\text{e-}03$, i.e., the difference between $\text{KLD}(p,\hat{p}')$ and $\text{KLD}(p,\hat{p}'')$ (and the one between $\text{KLD}(p,\hat{p}'')$ and $\text{KLD}(p,\hat{p}''')$) is 2 orders of magnitude, while the difference between $\text{KLD}(p,\hat{p}')$ and $\text{KLD}(p,\hat{p}''')$ is no less than 4 orders of magnitude. The increase in error (as computed by KLD) deriving from using $\hat{p}'''$ instead of $\hat{p}'$ is +632599%.

realistic quantifiers are (as we have observed above) very close to 0, for these values the $\text{NKLD}(p, \hat{p})$ curve is well approximated by $y = \frac{1}{2}\,\text{KLD}(p, \hat{p})$. As an EMQ, NKLD thus *de facto* inherits most of the problems of KLD .

All of the above shows that choosing a good EMQ (and the same may well be true for tasks other than quantification) should also be based, aside from the formal properties that the EMQ enjoys, on criteria that either resist or completely escape formalization, such as understandability and ease of use.

5.3 Evaluating Quantification across Multiple Samples

On a different note, we also need to stress a key difference between measures of classification accuracy and measures of quantification accuracy (or error). The objects of classification are individual unlabelled items, and all measures of classification accuracy (e.g., F_1) are defined with respect to a test *set* of such objects. The objects of quantification, instead, are samples, and all the measures of quantification accuracy we have discussed in this paper are defined on a *single* such sample (i.e., they measure how well the true distribution of the classes *across this individual sample* is approximated by the predicted distribution of the classes across the same sample). Since every evaluation is worthless if carried out on a single object, it is clear that quantification systems need to be evaluated on *sets* of samples. This means that every measure that we have discussed needs first to be evaluated on each sample, and then its global score across the test set (i.e., the set of samples on which testing is carried out) needs to be computed. This global score may be computed via any measure of central tendency, e.g., via an average, or a median, or other (for instance, if NAE is used, we might in turn use *Average* NAE or *Median* NAE, where averages and medians are computed across a set of samples). We do not take any specific stand for or against computing global scores via any specific measure of central tendency, since each of them may serve different but legitimate purposes. Note that a weighted average (in which the weight of a sample is inversely proportional to the score that the perverse estimator would obtain on the sample) might be appropriate for measures that do not satisfy **MAX**.

6 Conclusions

We have presented a study that "evaluates evaluation", in the tradition of the so-called "axiomatic" approach to the study of evaluation measures for information retrieval and related tasks. Our effort has targeted quantification, an important task at the crossroads of information retrieval, data mining, and machine learning, and has consisted of analysing previously proposed evaluation measures for quantification using the toolbox of the above-mentioned "axiomatic" approach. The work closest in spirit to the present one is our past work on the analysis of evaluation measures for classification (Sebastiani

2015). However, quantification poses more difficult problems than classification, since evaluation measures for quantification are inherently nonlinear (i.e., quantification error cannot be expressed as a linear function of the labelling error made on individual items). This is unlike classification, for which linear measures (e.g., standard accuracy, or K – see (Sebastiani 2015)) are possible.

We have proposed eight properties that, as we have argued, are desirable for measures that attempt to evaluate quantification (two such properties are actually mutually exclusive, and are desirable each in a different class of applications of quantification). Our analysis has revealed that, unfortunately, no existing evaluation measure for quantification satisfies all the other six properties. While this points to the fact that more research is needed to identify, or synthesize, a truly adequate such measure, this also means that, for the moment being, we have to evaluate the relative desirability of the properties that the existing measures do not satisfy. We have argued that some such properties are more important than others, and that as a result two measures ("Absolute Error" and "Relative Absolute Error") stand out as the most satisfactory ones (interestingly enough, they are also the most time-honoured ones, and the mathematically simplest ones).

As we have argued, RAE is more adequate for application contexts (e.g., quantifying the Tubercolosis class, as discussed in Section 3.1) in which an estimation error of a given absolute magnitude should be considered more serious if it affects a rare class, while AE is more adequate for those applications (e.g., quantifying the NoShow class, as discussed in Section 3.1) in which an estimation error of a given absolute magnitude has the same impact independently from the true prevalence of the affected class. Future work should also address the problem of how to best characterize these two classes of applications. The *number* and the *percentage* of items in a sample σ that belong to class c, seem to be essentially one and the same thing, but some applications (e.g., the NoShow example) are inherently interested in numbers, while other applications (e.g., the Tubercolosis example) seem more interested in percentages. When is it that a certain application belongs to the former (or to the latter) class, and why?

Aside from the design and use of an appropriate evaluation measure, there are further aspects concerning the evaluation of quantification that this work does not tackle. One of them is how to devise an evaluation *protocol* that strikes a balance between the two contrasting goals of (a) testing quantifiers on samples that exhibit *naturally occurring* class prevalences (this is the approach adopted in works such as (Gao and Sebastiani 2016; Nakov et al. 2016)), and (b) testing quantifiers also on samples that exhibit class prevalences (very) different from the naturally occurring ones (this is the approach adopted in works such as (Forman 2008; Esuli et al. 2018)). The realistic nature of the samples is the primary concern of the former approach, while testing quantifiers for robustness to different amounts of "prior probability shift" (i.e., difference between the prevalences in the training set and in the unlabelled set) is the one of the latter. We are working on an attempt to combine the strengths of both worlds, and hope to report results in the near future.

Acknowledgements

This work has benefitted from many discussions that I have had over the years with Andrea Esuli, Wei Gao, Ercan Kuruoglu, and Alejandro Moreo.

References

Rocío Alaíz-Rodríguez, Alicia Guerrero-Curieses, and Jesús Cid-Sueiro. 2011. Class and subclass probability re-estimation to adapt a classifier in the presence of concept drift. *Neurocomputing* 74, 16 (2011), 2614–2623. DOI:http://dx.doi.org/10.1016/j.neucom.2011.03.019

S. M. Ali and S. David Silvey. 1966. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B* 28, 1 (1966), 131–142. DOI:http://dx.doi.org/10.1111/j.2517-6161.1966.tb00626.x

Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. 2011. A comparison of evaluation metrics for document filtering. In *Proceedings of the 2nd International Conference of the Cross-Language Evaluation Forum (CLEF 2011)*. Amsterdam, NL, 38–49. DOI:http://dx.doi.org/10.1007/978-3-642-23708-9_6

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2013. Variable-constraint classification and quantification of radiology reports under the ACR Index. *Expert Systems and Applications* 40, 9 (2013), 3441–3449. DOI:http://dx.doi.org/10.1016/j.eswa.2012.12.052

José Barranquero, Jorge Díez, and Juan José del Coz. 2015. Quantification-oriented learning based on reliable classifiers. *Pattern Recognition* 48, 2 (2015), 591–604. DOI:http://dx.doi.org/10.1016/j.patcog.2014.07.032

José Barranquero, Pablo González, Jorge Díez, and Juan José del Coz. 2013. On the study of nearest neighbor algorithms for prevalence estimation in binary problems. *Pattern Recognition* 46, 2 (2013), 472–482. DOI:http://dx.doi.org/10.1016/j.patcog.2012.07.022

Oscar Beijbom, Judy Hoffman, Evan Yao, Trevor Darrell, Alberto Rodriguez-Ramirez, Manuel Gonzalez-Rivero, and Ove Hoegh-Guldberg. 2015. Quantification in-the-wild: Data-sets and baselines. (2015). CoRR abs/1510.04811 (2015). Presented at the NIPS 2015 Workshop on Transfer and Multi-Task Learning, Montreal, CA.

Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. 2010. Quantification via probability estimators. In *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM 2010)*. Sydney, AU, 737–742. DOI:http://dx.doi.org/10.1109/icdm.2010.75

Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. 2014. Aggregative quantification for regression. *Data Mining and Knowledge Discovery* 28, 2 (2014), 475–518. DOI:http://dx.doi.org/10.1007/s10618-013-0308-z

Luca Busin and Stefano Mizzaro. 2013. Axiometrics: An axiomatic approach to information retrieval effectiveness metrics. In *Proceedings of the 4th International Conference on the Theory of Information Retrieval (ICTIR 2013)*. Copenhagen, DK, 8. DOI:http://dx.doi.org/10.1145/2499178.2499182

Dallas Card and Noah A. Smith. 2018. The importance of calibration for estimating proportions from annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2018)*. New Orleans, US, 1636–1646. DOI:http://dx.doi.org/10.18653/v1/n18-1148

Andrea Ceron, Luigi Curini, and Stefano M. Iacus. 2016. iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. *Information Sciences* 367/368 (2016), 105—124. DOI:http://dx.doi.org/10.1016/j.ins.2016.05.052

Imre Csiszár and Paul C. Shields. 2004. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory* 1, 4 (2004), 417–528. DOI:http://dx.doi.org/10.1561/0100000004

Giovanni Da San Martino, Wei Gao, and Fabrizio Sebastiani. 2016a. Ordinal text quantification. In *Proceedings of the 39th ACM Conference on Research and Development in Information Retrieval (SIGIR 2016)*. Pisa, IT, 937–940. DOI:http://dx.doi.org/10.1145/2911451.2914749

Giovanni Da San Martino, Wei Gao, and Fabrizio Sebastiani. 2016b. QCRI at SemEval-2016 Task 4: Probabilistic methods for binary and ordinal quantification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. San Diego, US, 58–63. DOI:http://dx.doi.org/10.18653/v1/s16-1006

Marthinus C. du Plessis, Gang Niu, and Masashi Sugiyama. 2017. Class-prior estimation for learning from positive and unlabeled data. *Machine Learning* 106, 4 (2017), 463–492. DOI:http://dx.doi.org/10.1007/s10994-016-5604-6

Marthinus C. du Plessis and Masashi Sugiyama. 2012. Semi-supervised learning of class balance under class-prior change by distribution matching. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*. Edinburgh, UK.

Marthinus C. du Plessis and Masashi Sugiyama. 2014. Class prior estimation from positive and unlabeled data. *IEICE Transactions* 97-D, 5 (2014), 1358–1362. DOI:http://dx.doi.org/10.1587/transinf.e97.d.1358

Andrea Esuli. 2016. ISTI-CNR at SemEval-2016 Task 4: Quantification on an ordinal scale. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. San Diego, US. DOI:http://dx.doi.org/10.18653/v1/s16-1011

Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. 2018. A recurrent neural network for sentiment quantification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018)*. Torino, IT, 1775–1778. DOI:http://dx.doi.org/10.1145/3269206.3269287

Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. 2019. Cross-lingual sentiment quantification. (2019). arXiv:1904.07965.

Andrea Esuli and Fabrizio Sebastiani. 2010. Sentiment quantification. *IEEE Intelligent Systems* 25, 4 (2010), 72–75.

Andrea Esuli and Fabrizio Sebastiani. 2014. Explicit loss minimization in quantification applications (preliminary draft). In *Proceedings of the 8th International Workshop on Information Filtering and Retrieval (DART 2014)*. Pisa, IT, 1–11.

Andrea Esuli and Fabrizio Sebastiani. 2015. Optimizing text quantifiers for multivariate loss functions. *ACM Transactions on Knowledge Discovery and Data* 9, 4 (2015), Article 27. DOI:http://dx.doi.org/10.1145/2700406

Afonso Fernandes Vaz, Rafael Izbicki, and Rafael Bassi Stern. 2019. Quantification under prior probability shift: The ratio estimator and its extensions. *Journal of Machine Learning Research* 20 (2019), 79:1–79:33.

Marco Ferrante, Nicola Ferro, and Maria Maistro. 2015. Towards a formal framework for utility-oriented measurements of retrieval effectiveness. In *Proceedings of the 5th ACM International Conference on the Theory of Information Retrieval (ICTIR 2015)*. Northampton, US, 21–30. DOI:http://dx.doi.org/10.1145/2808194.2809452

Marco Ferrante, Nicola Ferro, and Silvia Pontarollo. 2018. A general theory of IR evaluation measures. *IEEE Transactions on Knowledge and Data Engineering* (2018). DOI:http://dx.doi.org/10.1109/TKDE.2018.2840708

George Forman. 2005. Counting positives accurately despite inaccurate classification. In *Proceedings of the 16th European Conference on Machine Learning (ECML 2005)*. Porto, PT, 564–575. DOI:http://dx.doi.org/10.1007/11564096_55

George Forman. 2006. Quantifying trends accurately despite classifier error and class imbalance. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*. Philadelphia, US, 157–166. DOI:http://dx.doi.org/10.1145/1150402.1150423

George Forman. 2008. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery* 17, 2 (2008), 164–206. DOI:http://dx.doi.org/10.1007/s10618-008-0097-y

Wei Gao and Fabrizio Sebastiani. 2015. Tweet sentiment: From classification to quantification. In *Proceedings of the 7th International Conference on Advances in Social Network Analysis and Mining (ASONAM 2015)*. Paris, FR, 97–104. DOI:http://dx.doi.org/10.1145/2808797.2809327

Wei Gao and Fabrizio Sebastiani. 2016. From classification to quantification in tweet sentiment analysis. *Social Network Analysis and Mining* 6, 19 (2016), 1–22. DOI: http://dx.doi.org/10.1007/s13278-016-0327-z

Pablo González, Eva Álvarez, Jorge Díez, Ángel López-Urrutia, and Juan J. del Coz. 2017. Validation methods for plankton image classification systems. *Limnology and Oceanography: Methods* 15 (2017), 221–237. DOI:http://dx.doi.org/10.1002/lom3.10151

Pablo González, Alberto Castaño, Nitesh V. Chawla, and Juan José del Coz. 2017. A review on quantification learning. *Comput. Surveys* 50, 5 (2017), 74:1–74:40. DOI: http://dx.doi.org/10.1145/3117807

Pablo González, Jorge Díez, Nitesh Chawla, and Juan José del Coz. 2017. Why is quantification an interesting learning problem? *Progress in Artificial Intelligence* 6, 1 (2017), 53–58. DOI:http://dx.doi.org/10.1007/s13748-016-0103-3

Víctor González-Castro, Rocío Alaiz-Rodríguez, and Enrique Alegre. 2013. Class distribution estimation based on the Hellinger distance. *Information Sciences* 218 (2013), 146–164. DOI:http://dx.doi.org/10.1016/j.ins.2012.05.028

Víctor González-Castro, Rocío Alaiz-Rodríguez, Laura Fernández-Robles, R. Guzmán-Martínez, and Enrique Alegre. 2010. Estimating class proportions in boar semen analysis using the Hellinger distance. In *Proceedings of the 23rd International Conference on Industrial Engineering and other Applications of Applied Intelligent Systems (IEA/AIE 2010)*. Cordoba, ES, 284–293. DOI:http://dx.doi.org/10.1007/978-3-642-13022-9_29

Daniel J. Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54, 1 (2010), 229–247. DOI:http://dx.doi.org/10.1111/j.1540-5907.2009.00428.x

Purushottam Kar, Shuai Li, Harikrishna Narasimhan, Sanjay Chawla, and Fabrizio Sebastiani. 2016. Online optimization methods for the quantification problem. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*. San Francisco, US, 1625–1634. DOI:http://dx.doi.org/10.1145/2939672.2939832

Katherine A. Keith and Brendan O'Connor. 2018. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*. Brussels, BE.

Gary King and Ying Lu. 2008. Verbal autopsy methods with multiple causes of death. *Statist. Sci.* 23, 1 (2008), 78–91. DOI:http://dx.doi.org/10.1214/07-sts247

Roy Levin and Haggai Roitman. 2017. Enhanced probabilistic classify and count methods for multi-label text quantification. In *Proceedings of the 7th ACM International Conference on the Theory of Information Retrieval (ICTIR 2017)*. Amsterdam, NL, 229–232. DOI: http://dx.doi.org/10.1145/3121050.3121083

Friedrich Liese and Igor Vajda. 2006. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory* 52, 10 (2006), 4394–4412. DOI:http://dx.doi.org/10.1109/tit.2006.881731

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151. DOI:http://dx.doi.org/10.1109/18.61115

David J. MacKay. 2003. *Information theory, inference and learning algorithms*. Cambridge University Press, Cambridge, UK.

André G. Maletzke, Denis Moreira dos Reis, and Gustavo E. Batista. 2017. Quantification in data streams: Initial results. In *Proceedings of the 2017 Brazilian Conference on Intelligent Systems (BRACIS 2017)*. Uberlândia, BZ, 43–48. DOI:http://dx.doi.org/10.1109/BRACIS.2017.74

André G. Maletzke, Denis Moreira dos Reis, and Gustavo E. Batista. 2018. Combining instance selection and self-training to improve data stream quantification. *Journal of the Brazilian Computer Society* 24, 12 (2018), 43–48. DOI:http://dx.doi.org/10.1186/s13173-018-0076-0

Letizia Milli, Anna Monreale, Giulio Rossetti, Fosca Giannotti, Dino Pedreschi, and Fabrizio Sebastiani. 2013. Quantification trees. In *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM 2013)*. Dallas, US, 528–536. DOI:http://dx.doi.org/10.1109/icdm.2013.122

Letizia Milli, Anna Monreale, Giulio Rossetti, Dino Pedreschi, Fosca Giannotti, and Fabrizio Sebastiani. 2015. Quantification in social networks. In *Proceedings of the 2nd IEEE International Conference on Data Science and Advanced Analytics (DSAA 2015)*. Paris, FR. DOI:http://dx.doi.org/10.1109/dsaa.2015.7344845

Alastair Moffat. 2013. Seven numeric properties of effectiveness metrics. In *Proceedings of the 9th Conference of the Asia Information Retrieval Societies (AIRS 2013)*. Singapore, SN, 1–12. DOI:http://dx.doi.org/10.1007/978-3-642-45068-6_1

Denis Moreira dos Reis, André Maletzke, Everton Cherman, and Gustavo E. Batista. 2018a. One-class quantification. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2018)*. Dublin, IE.

Denis Moreira dos Reis, André G. Maletzke, Diego F. Silva, and Gustavo E. Batista. 2018b. Classifying and counting with recurrent contexts. In *Proceedings of the 24th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2018)*. London, UK, 1983–1992. DOI:http://dx.doi.org/10.1145/3219819.3220059

Preslav Nakov, Noura Farra, and Sara Rosenthal. 2017. SemEval-2017 Task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, CA. DOI:http://dx.doi.org/10.18653/v1/s17-2088

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 Task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. San Diego, US, 1–18. DOI:http://dx.doi.org/10.18653/v1/s16-1001

Pablo Pérez-Gállego, Alberto Castaño, José Ramón Quevedo, and Juan José del Coz. 2019. Dynamic ensemble selection for quantification tasks. *Information Fusion* 45 (2019), 1–15. DOI:http://dx.doi.org/10.1016/j.inffus.2018.01.001

Pablo Pérez-Gállego, José Ramón Quevedo, and Juan José del Coz. 2017. Using ensembles for problems with characterizable changes in data distribution: A case study on quantification. *Information Fusion* 34 (2017), 87–100. DOI:http://dx.doi.org/10.1016/j.inffus.2016.07.001

Marco Saerens, Patrice Latinne, and Christine Decaestecker. 2002. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation* 14, 1 (2002), 21–41. DOI:http://dx.doi.org/10.1162/089976602753284446

Amartya Sanya, Pawan Kumar, Purushottam Kar, Sanjay Chawla, and Fabrizio Sebastiani. 2018. Optimizing non-decomposable measures with deep networks. *Machine Learning* 107, 8-10 (2018), 1597–1620. DOI:http://dx.doi.org/10.1007/s10994-018-5736-y

Fabrizio Sebastiani. 2015. An axiomatically derived measure for the evaluation of classification algorithms. In *Proceedings of the 5th ACM International Conference on the Theory of Information Retrieval (ICTIR 2015)*. Northampton, US, 11–20. DOI:http://dx.doi.org/10.1145/2808194.2809449

David J. Spence. 2019. *Quantification under class-conditional dataset shift*. Ph.D. Dissertation. University of Sussex, Brighton, UK.

Lei Tang, Huiji Gao, and Huan Liu. 2010. Network quantification despite biased labels. In *Proceedings of the 8th Workshop on Mining and Learning with Graphs (MLG 2010)*. Washington, US, 147–154. DOI:http://dx.doi.org/10.1145/1830252.1830271

Dirk Tasche. 2017. Fisher consistency for prior probability shift. *Journal of Machine Learning Research* 18 (2017), 95:1–95:32.

Zhihao Zhang and Jie Zhou. 2010. Transfer estimation of evolving class priors in data stream classification. *Pattern Recognition* 43, 9 (2010), 3151–3161. DOI:http://dx.doi.org/j.patcog.2010.03.021

A Properties of AE

We here prove that AE enjoys **IoI**, **NN**, **IND**, **MON**, **IMP**, **ABS**. While some of these proofs are trivial, these are reported in detail in order to show how the same arguments can be used to prove the same for many of the other EMQs discussed in Section 4.

AE enjoys **IoI**. In fact, $\mathrm{AE}(p,\hat{p}) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}|\hat{p}(c)-p(c)| = 0$ implies that $\sum_{c\in\mathcal{C}}|\hat{p}(c)-p(c)| = 0$; given that $\sum_{c\in\mathcal{C}}|\hat{p}(c)-p(c)|$ is a sum of nonnegative factors, this implies that $|\hat{p}(c)-p(c)| = 0$ for all $c \in \mathcal{C}$, i.e., $\hat{p}(c) = p(c)$ for all $c \in \mathcal{C}$. Conversely, if $\hat{p} = p$, then $\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}|\hat{p}(c)-p(c)| = 0$. □

AE enjoys **NN**. Quite obviously, $\frac{1}{|\mathcal{C}|} \geq 0$ and $\sum_{c\in\mathcal{C}}|\hat{p}(c)-p(c)| \geq 0$, which implies that $\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}|\hat{p}(c)-p(c)| \geq 0$. □

AE enjoys **IND**. Given codeframe $\mathcal{C} = \{c_1, ..., c_k, c_{k+1}, ..., c_n\}$, for any true distribution p on $\mathcal{C}$ and predicted distributions $\hat{p}'$ and $\hat{p}''$ on $\mathcal{C}$ such that $\hat{p}'(c) = \hat{p}''(c)$ for all $c \in \{c_{k+1}, ..., c_n\}$, the inequality

$$\mathrm{AE}(p,\hat{p}') \leq \mathrm{AE}(p,\hat{p}'')$$

resolves to

$$\begin{aligned}
\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}|\hat{p}'(c)-p(c)| &\leq \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}|\hat{p}''(c)-p(c)| \\
\sum_{c\in\mathcal{C}}|\hat{p}'(c)-p(c)| &\leq \sum_{c\in\mathcal{C}}|\hat{p}''(c)-p(c)| \\
\sum_{c\in\mathcal{C}_1}|\hat{p}'(c)-p(c)| + \sum_{c\in\mathcal{C}_2}|\hat{p}'(c)-p(c)| &\leq \sum_{c\in\mathcal{C}_1}|\hat{p}''(c)-p(c)| + \sum_{c\in\mathcal{C}_2}|\hat{p}''(c)-p(c)| \\
\sum_{c\in\mathcal{C}_1}|\hat{p}'(c)-p(c)| &\leq \sum_{c\in\mathcal{C}_1}|\hat{p}''(c)-p(c)| \\
\frac{1}{|\mathcal{C}_1|}\sum_{c\in\mathcal{C}_1}|\hat{p}'(c)-p(c)| &\leq \frac{1}{|\mathcal{C}_1|}\sum_{c\in\mathcal{C}_1}|\hat{p}''(c)-p(c)| \\
\mathrm{AE}(p_{\mathcal{C}_1},\hat{p}'_{\mathcal{C}_1}) &\leq \mathrm{AE}(p_{\mathcal{C}_1},\hat{p}''_{\mathcal{C}_1}) \qquad \square
\end{aligned}$$

AE enjoys **MON**. This can be proven by showing that AE enjoys **B-MON**, since we have proven that it enjoys **IND**. Given codeframe $\mathcal{C} = \{c_1, c_2\}$ and true distribution p, if predicted distributions $\hat{p}', \hat{p}''$ are such that $\hat{p}''(c_1) < \hat{p}'(c_1) \leq p(c_1)$, then it holds that

$$\begin{aligned}
\mathrm{AE}(p,\hat{p}') &= \frac{1}{2}(|\hat{p}'(c_1)-p(c_1)| + |\hat{p}'(c_2)-p(c_2)|) \\
&< \frac{1}{2}(|\hat{p}''(c_1)-p(c_1)| + |\hat{p}''(c_2)-p(c_2)|) \\
&= \mathrm{AE}(p,\hat{p}'') \qquad \square
\end{aligned}$$

AE enjoys **IMP**. This can be shown by showing that AE enjoys **B-IMP**, since we have proven that it enjoys **IND**. Given codeframe $\mathcal{C} = \{c_1, c_2\}$, true distribution p, predicted distributions $\hat{p}'$ and $\hat{p}''$, and constant $a \geq 0$ such that $\hat{p}'(c_1) = p(c_1) + a$ and $\hat{p}''(c_1) = p(c_1) - a$, it holds that

$$\begin{aligned}
\mathrm{AE}(p,\hat{p}') &= \frac{1}{2}(|\hat{p}'(c_1)-p(c_1)| + |\hat{p}'(c_2)-p(c_2)|) \\
&= \frac{1}{2}(|(p(c_1)+a)-p(c_1)| + |(p(c_2)-a)-p(c_2)|) \\
&= \frac{1}{2}(|a| + |-a|) \\
&= \frac{1}{2}(|-a| + |a|) \\
&= \frac{1}{2}(|(p(c_1)-a)-p(c_1)| + |(p(c_2)+a)-p(c_2)|) \\
&= \mathrm{AE}(p,\hat{p}'') \qquad \square
\end{aligned}$$

AE enjoys **ABS**. This can be shown by showing that AE enjoys **B-ABS**, since we have proven that it enjoys **IND**. Given codeframe $\mathcal{C} = \{c_1, c_2\}$, constant $a > 0$, true distributions

p' and p'' such that $p'(c_1) < p''(c_1)$ and $p''(c_1) < p''(c_2)$, if a predicted distribution $\hat{p}'$ that estimates p' is such that $\hat{p}'(c_1) = p'(c_1) \pm a$ and a predicted distribution $\hat{p}''$ that estimates p'' is such that $\hat{p}''(c_1) = p''(c_1) \pm a$, then it holds that

$$\begin{aligned}
\mathrm{AE}(p', \hat{p}') &= \frac{1}{2}(|\hat{p}'(c_1) - p'(c_1)| + |\hat{p}'(c_2) - p'(c_2)|) \\
&= \frac{1}{2}(|(p'(c_1) \pm a) - p'(c_1)| + |(p'(c_2) \mp a) - p'(c_2)|) \\
&= \frac{1}{2}(2a) \\
&= \frac{1}{2}((p''(c_1) \pm a) - p''(c_1)| + |(p''(c_2) \mp a) - p''(c_2)|) \\
&= \frac{1}{2}(|\hat{p}''(c_1) - p''(c_1)| + |\hat{p}''(c_2) - p''(c_2)|) \\
&= \mathrm{AE}(p'', \hat{p}'') \qquad \square
\end{aligned}$$

B Testing for MAX, IMP, ABS, REL

In this section we present simple tests aimed at establishing that a certain EMQ D does *not* enjoy a certain property $\pi \in \{\mathbf{MAX}, \mathbf{IMP}, \mathbf{ABS}, \mathbf{REL}\}$. The basic pattern of these tests is to show that π does not hold for D by providing a counterexample. More in particular, given a concrete scenario s characterized by (1) a codeframe $\mathcal{C}$, (2) one or more true distributions p_1, p_2, ..., and (3) one or more predicted distributions $\hat{p}_1$, $\hat{p}_2$, ..., the test attempts to check whether the scenario satisfies the particular constraint that is required for property π to hold. Since for D to enjoy property π the constraint is required to hold for all scenarios, if π does not hold in s we can conclude than D does not enjoy π. Instead, if π does hold in s we can conclude nothing, and thus need to study the issue further.

B.1 A Counterexample for MAX

In the test for **MAX** we consider the scenario described in the following table

	$p(c_1)$	$p(c_2)$	$\hat{p}(c_1)$	$\hat{p}(c_2)$	AE	NAE	RAE	NRAE	SE	DR	KLD	NKLD	PD
p'	0.01	0.99	1.00	0.00	0.9900	1.0000	49.9975	1.0000	0.9801	0.9950	14.3076	0.9999	980100.0004
p''	0.49	0.51	1.00	0.00	0.5100	1.0000	1.0204	1.0000	0.2601	0.7550	6.7065	0.9975	260100.0001

and characterized by two different true distributions (1st and 2nd row) across the same codeframe $\mathcal{C} = \{c_1, c_2\}$.[19] The test consists in checking whether their respective perverse estimators obtain from D the same score: if the values of measure D in the two rows are not the same (greyed-out cells), this implies that D does not satisfy **MAX** (if they are the same, this does *not* necessarily mean that D satisfies **MAX**). Concerning the values obtained by NKLD, see the discussion in Footnote 14.

The table shows that none of AE, RAE, SE, DR, KLD, PD satisfies **MAX**.

[19] We assume $|D| = 1,000,000$. This assumption has no relevance on the qualitative conclusions we draw here, and only affects the magnitude of the values in the table (since the value of $|D|$ affects the value of ϵ, and thus of RAE, NRAE, DR, KLD, NKLD, PD – see Section 4.3) and following.

B.2 A Counterexample for **IMP**

In the test for **IMP** we consider the scenario described in the following table

	$p(c_1)$	$p(c_2)$	$\hat{p}(c_1)$	$\hat{p}(c_2)$	AE	NAE	RAE	NRAE	SE	DR	KLD	NKLD	PD
p'	0.20	0.80	0.25	0.75	0.0500	0.0625	0.1562	0.0625	0.0025	0.1312	0.0070	0.0035	0.0117
p''	0.20	0.80	0.15	0.85	0.0500	0.0625	0.1562	0.0625	0.0025	0.1544	0.0090	0.0045	0.0181

and characterized by a codeframe $\mathcal{C} = \{c_1, c_2\}$, a true distribution p (Columns 2 and 3), and two predicted distributions $\hat{p}'$ and $\hat{p}''$ (Columns 4 and 5, Rows 2 and 3) which are such that (i) $\hat{p}'$ overestimates and $\hat{p}''$ underestimates the prevalence of a class c_1 by a certain amount $a > 0$ (here: 0.05), and, symmetrically, (ii) $\hat{p}'$ underestimates and $\hat{p}''$ overestimates the prevalence of another class c_2 by the same amount a. If the values of $D(p, \hat{p}')$ and $D(p, \hat{p}'')$ are not the same (which in the table is indicated by greyed-out cells), this implies that D does not satisfy **IMP** (if they are the same, this does *not* necessarily mean that D satisfies **IMP**).

The table shows that none of DR, KLD, NKLD, PD satisfies **IMP**.

B.3 A Counterexample for **REL**

In the test for **REL** we consider the scenario described in the following table

	$p(c_1)$	$p(c_2)$	$\hat{p}(c_1)$	$\hat{p}(c_2)$	AE	NAE	RAE	NRAE	SE	DR	KLD	NKLD	PD
p'	0.20	0.80	0.70	0.30	0.5000	0.6250	1.5625	0.6250	0.2500	0.6696	0.5341	0.2609	0.7738
p''	0.25	0.75	0.75	0.25	0.5000	0.6667	1.3333	0.6667	0.2500	0.6667	0.5493	0.2679	0.8333

with a codeframe $\mathcal{C} = \{c_1, c_2\}$, two true distributions p' and p'' (Rows 2 and 3, Columns 2 to 4), and two corresponding predicted distributions $\hat{p}'$ and $\hat{p}''$ (Rows 2 and 3, Columns 5 to 7), such that in both cases the predicted distribution overestimates the prevalence of c_1 by the same amount $a > 0$ (here: 0.50), with $p'(c_1) < p''(c_1)$. Here, if it is not the case that $D(p, \hat{p}') > D(p, \hat{p}'')$ (which in the table is indicated by greyed-out cells), then D does not satisfy **REL** (if $D(p, \hat{p}') \neq D(p, \hat{p}'')$, this does *not* necessarily mean that D satisfies **REL**).

The table shows that none of AE, NAE, NRAE, SE, KLD, NKLD, PD satisfies **REL**.

B.4 A Counterexample for **ABS**

In the test for **ABS** we consider the same scenario as described in Appendix B.3, i.e.,

	$p(c_1)$	$p(c_2)$	$\hat{p}(c_1)$	$\hat{p}(c_2)$	AE	NAE	RAE	NRAE	SE	DR	KLD	NKLD	PD
p'	0.20	0.80	0.70	0.30	0.5000	0.6250	1.5625	0.6250	0.2500	0.6696	0.5341	0.2609	0.7738
p''	0.25	0.75	0.75	0.25	0.5000	0.6667	1.3333	0.6667	0.2500	0.6667	0.5493	0.2679	0.8333

with a codeframe $\mathcal{C} = \{c_1, c_2\}$, two true distributions p' and p'' (Rows 2 and 3, Columns 2 to 4), and two corresponding predicted distributions $\hat{p}'$ and $\hat{p}''$ (Rows 2 and 3, Columns 5 to 7), such that in both cases the predicted distribution overestimates the prevalence of c_1 by the same amount $a > 0$ (here: 0.50), with $p'(c_1) < p''(c_1)$. Here, if the values of $D(p, \hat{p}')$ and $D(p, \hat{p}'')$ are not equal (which in the table is indicated by greyed-out cells), this implies that D does not satisfy **ABS** (if $D(p, \hat{p}') = D(p, \hat{p}'')$, this does *not* necessarily mean that D satisfies **ABS**).

The table shows that none of NAE, RAE, NRAE, DR, KLD, NKLD, PD satisfies **ABS**.

C Proving that MON Holds

In this section we prove that **MON** holds for KLD and PD. For this it will be sufficient to prove that KLD and PD enjoy **B-MON**, since it is immediate to verify that KLD and PD enjoy **IND**.

For ease of exposition, let us define the shorthands $a \equiv p(c_1)$ and $x \equiv \hat{p}(c_1)$.[20] In order to show that D satisfies **B-MON** it is sufficient to show that

1. if $(a-x) > 0$, then $\frac{\partial}{\partial(a-x)} D > 0$ for $a, x, (a-x) \in (0,1)$
2. if $(x-a) > 0$, then $\frac{\partial}{\partial(x-a)} D > 0$ for $a, x, (x-a) \in (0,1)$

because an increase in $(a-x) = (p(c_1)-\hat{p}(c_1))$ implies an equivalent increase in $(p(c_2)-\hat{p}(c_2))$ (same for $(x-a)$).

Theorem 1 KLD *satisfies* ***B-MON****.*

Proof. We first treat the case $(a-x) > 0$; let us define $y \equiv (a-x)$. In this case

$$\begin{aligned}
\frac{\partial}{\partial y}\,\mathrm{KLD} &= \frac{\partial}{\partial y}(a \log \frac{a}{x} + (1-a)\log\frac{1-a}{1-x}) \\
&= \frac{\partial}{\partial y}(a \log \frac{a}{a-y} + (1-a)\log\frac{1-a}{1-a+y}) \\
&= \frac{-y}{(a-y-1)(a-y)} \\
&= \frac{x-a}{(x-1)x}
\end{aligned}$$

Since we are in the case in which $(x-a) < 0$, and since $(x-1) < 0$ and $x > 0$, then $\frac{x-a}{(x-1)x} > 0$ for all $a, x, (a-x) \in (0,1)$.

Let us now treat the case $(x-a) > 0$, and let us define $y \equiv (x-a)$. In this case

$$\begin{aligned}
\frac{\partial}{\partial y}\,\mathrm{KLD} &= \frac{\partial}{\partial y}(a \log \frac{a}{x} + (1-a)\log\frac{1-a}{1-x}) \\
&= \frac{\partial}{\partial y}(a \log \frac{a}{y-a} + (1-a)\log\frac{1-a}{1-y+a}) \\
&= \frac{-y}{(a+y-1)(a+y)} \\
&= \frac{a-x}{(x-1)x}
\end{aligned}$$

Since in this case it holds that $(a-x) < 0$, and since $(x-1) < 0$ and $x > 0$, then $\frac{x-a}{(x-1)x} > 0$ for all $a, x, (x-a) \in (0,1)$. This concludes our proof. □

Theorem 2 PD *satisfies* ***B-MON****.*

[20] For the EMQs that require smoothed probabilities to be used, these definitions obviously need to be replaced by $a \equiv p_s(c_1)$ and $x \equiv \hat{p}_s(c_1)$.

PROOF. We first treat the case $(a - x) > 0$; let us define $y \equiv (a - x)$. In this case

$$\begin{aligned}\frac{\partial}{\partial y}\,\mathrm{PD} &= \frac{\partial}{\partial y}\left(\frac{(a-x)^2}{x} + \frac{((1-a)-(1-x))^2}{1-x}\right)\\ &= \frac{\partial}{\partial y}\left(\frac{y^2}{a-y} + \frac{((1-a)-(1-a+y))^2}{1-a+y}\right)\\ &= \frac{y(-2a^2+2a(y+1)-y)}{(a-y)^2(-a+y+1)^2}\\ &= \frac{(a-x)(a-2ax+x)}{x^2(1-x)^2}\end{aligned}$$

Since in this case it holds that $a > x$, it is true that that $(a - 2ax + x) > (x - 2ax + x) = 2x(1 - a) > 0$, since by hypothesis it holds that $x, a \in (0, 1)$. Therefore, $\frac{\partial}{\partial y}\,\mathrm{PD} = \frac{(a-x)(a-2ax+x)}{x^2(1-x)^2} > 0$, since the two factors at the numerator and the two factors at the denominator are all strictly > 0.

Let us now treat the case $(x - a) > 0$, and let us define $y \equiv (x - a)$. In this case

$$\begin{aligned}\frac{\partial}{\partial y}\,\mathrm{PD} &= \frac{\partial}{\partial y}\left(\frac{(a-x)^2}{x} + \frac{((1-a)-(1-x))^2}{1-x}\right)\\ &= \frac{\partial\left(\frac{y^2}{y+a} + \frac{((1-a)-(1-a-y))^2}{1-a-y}\right)}{\partial y}\\ &= \frac{y(-2a^2-2a(y-1)+y)}{(a-y+1)^2(a+y)^2}\\ &= \frac{(x-a)(-2ax+x+a)}{(2a-x+1)^2x^2}\end{aligned}$$

Since in this case it holds that $x > a$, it is true that that $(-2ax + x + a) > (-2ax + 2a) = 2a(1 - x) > 0$, since by hypothesis it holds that $x, a \in (0, 1)$. Therefore, $\frac{\partial\,\mathrm{PD}}{\partial y} = \frac{(x-a)(-2ax+x+a)}{(2a-x+1)^2x^2} > 0$, since the two factors at the numerator and the two factors at the denominator are all strictly > 0. This concludes our proof. □

Table 2 Research works about quantification where the EMQs discussed in this paper have been first proposed (★) and later used (✓).

	AE	NAE	RAE	NRAE	SE	DR	KLD	NKLD	PD
(Saerens et al. 2002)	★								
(Forman 2005)	✓						★		
(Forman 2006)	✓						✓		
(Forman 2008)	✓						✓		
(Tang et al. 2010)	✓						✓		
(Bella et al. 2010)	✓				★				
(González-Castro et al. 2010)	✓		★						
(Zhang and Zhou 2010)	✓								
(Alaíz-Rodríguez et al. 2011)	✓		✓						
(Milli et al. 2013)							✓		
(Barranquero et al. 2013)	✓								
(González-Castro et al. 2013)	✓		✓						
(Esuli and Sebastiani 2014)	✓	★	✓	★			✓	★	
(du Plessis and Sugiyama 2014)					✓				
(Esuli and Sebastiani 2015)			✓				✓		
(Gao and Sebastiani 2015)	✓	✓	✓	✓			✓	✓	
(Barranquero et al. 2015)	✓						✓		
(Beijbom et al. 2015)	✓								
(Milli et al. 2015)							✓		
(Gao and Sebastiani 2016)	✓	✓	✓	✓			✓	✓	
(Ceron et al. 2016)	✓								★
(Kar et al. 2016)							✓		
(Nakov et al. 2016)							✓		
(González et al. 2017)	✓								
(du Plessis et al. 2017)					✓				
(Levin and Roitman 2017)						★			
(Pérez-Gállego et al. 2017)	✓				✓				
(Tasche 2017)			✓						
(Nakov et al. 2017)	✓		✓				✓		
(Maletzke et al. 2017)	✓	✓		✓					
(Maletzke et al. 2018)				✓					
(Esuli et al. 2018)	✓		✓				✓		
(Card and Smith 2018)	✓								
(Moreira dos Reis et al. 2018a)	✓								
(Moreira dos Reis et al. 2018b)	✓			✓					
(Sanya et al. 2018)							✓		
(Keith and O'Connor 2018)	✓								
(Fernandes Vaz et al. 2019)					✓				
(Esuli et al. 2019)	✓		✓				✓		
(Pérez-Gállego et al. 2019)	✓				✓				
(Spence 2019)	✓								