




On the nature of information access evaluation metrics: a unifying framework

Enrique Amigó¹ · Stefano Mizzaro² 

Received: 21 June 2019 / Accepted: 5 May 2020 / Published online: 29 May 2020
© Springer Nature B.V. 2020

Abstract

We provide a uniform, general, and complete formal account of evaluation metrics for ranking, classification, clustering, and other information access problems. We leverage concepts from measurement theory, such as scale types and permissible transformation functions, and we capture the nature of evaluation metrics in many tasks by two formal definitions, which lead to a distinction of two metric/tasks families, and provide a comprehensive classification of the tasks that have been proposed so far. We derive some theorems to analyze the suitability (or otherwise) of some common metrics. Within our model we can derive and explain the theoretical properties and drawbacks of the state of the art metrics for multiple tasks. The main contributions of this paper are that, differently from previous studies, the formalization is well grounded on a solid discipline, it is general as it can take into account most effectiveness metrics as well as most existing tasks, and it allows to derive important consequences on metrics and their limitations.

Keywords Evaluation · Measurement theory · Effectiveness · Accuracy · Metrics

1 Introduction

The number of evaluation effectiveness metrics in information access tasks is very large, and growing: in Information Retrieval (IR) alone, more than 100 effectiveness metrics exist, not taking into account the user oriented and Web-oriented ones (Amigó et al. 2014); in clustering various accuracy measures are used, even in official experimental initiatives and sometimes with undesirable properties (Amigó et al. 2009); in filtering the situation is analogous (Amigó et al. 2011); and of course, when considering other tasks the number of used metrics grows even more.

Being the evaluation scenario so rich and complex, it is not surprising that attempts have been made to understand, model, and formalize it. More in detail, researchers have defined

✉ Stefano Mizzaro
mizzaro@uniud.it

Enrique Amigó
enrique@lsi.uned.es

¹ UNED, Madrid, Spain

² Department of Mathematics, Computer Science, and Physics, University of Udine, Udine, Italy

formal properties (or axioms, or constraints) that must be satisfied by metrics. This has happened both in the early years (Van Rijsbergen 1974; Bollmann 1984) and more recently, with a renewed interest and several studies published in the last 5 years or so (Amigó et al. 2009, 2011, 2013, 2014, 2015; Moffat 2013; Busin and Mizzaro 2013; Maddalena and Mizzaro 2014; Ferrante et al. 2015; Sebastiani 2015; Ferrante et al. 2017). All these studies have in common a formal attitude: to try to understand in a formal way properties of effectiveness metrics. This paper follows the path of these studies, and addresses the formalization of effectiveness evaluation.

As it will be detailed in the following, this paper differs from previous studies in several respects. First, we aim at a more general approach: we do not focus on a specific task only, as others have done, but we take into account several information access tasks at the same time and we provide a uniform account. In this respect, one issue is that the number of tasks in information access is very large, and their variety is quite high: researchers build systems, for example, to classify tweets, cluster terms, retrieve documents, filter news, recommend movies, summarize texts, etc. To be able to provide a general and systematic account, we make two choices. First, we focus on the tasks that can be modeled by the assignment of a value to each item; most of the above mentioned examples match this description (the only exception being summarization). Second, we make an abstraction effort and we distill these many existing tasks into just four: *Classification* (assigning a category to each document), *Clustering* (organising documents into groups), *Ranking* (sorting documents), and *Quantitation* (assigning a numeric value to each document). In an attempt to avoid confusion, we try to use a specific terminology: we use the term *abstract task* to refer to the latter four only (i.e., Classification, Clustering, Ranking, and Quantitation) and we use the term *task* to refer to all the former ones. Although the terminology is somehow different from what commonly used in the literature, we believe that the small effort is worthwhile, and it allows us to precisely define the four abstract tasks we are addressing.

A second difference from previous studies is that we ground on measurement theory, that provides several useful notions including the measurement scale. We are not by any means the first ones to use measurement theory as a tool. However, we use it in a way that is different from previous work, and that is important to state upfront also to avoid confusion. It is probably intuitive to directly link the notion of metric with measurement: a metric *measures* system accuracy/effectiveness. This is the approach of the seminal work by Van Rijsbergen (1974) and also of more recent proposals by Ferrante et al. (2017, 2019). However, we do not do so, and in our approach measurement theory is used in a different way: our starting point is the fact that *both system outputs and gold standards can be seen as assignments of values to documents*, for all four abstract tasks. For example, the output of a classification system can be seen as an assignment of values on the nominal scale type; a rank of documents (a typical search engine output) can be seen as an assignment of values to documents, to be interpreted on the ordinal scale type (i.e., considering only the rank induced by the assigned values); and so on. Thus, measurement theory allows us to model both system outputs and gold standards as assignments of values as well as to state a direct relationship between the abstract tasks and the scale types.

Then, coming to the third difference from previous studies, in our framework an evaluation metric *compares two assignments* of numbers to documents: one provided by a system, and another one provided by human assessments. We define the effectiveness/accuracy of the system as the *closeness* of the two assignments. In other terms, measurement theory allows us to distinguish the above four abstract tasks on the basis of the scale used when assigning the values. For example, the nominal scale is used for classification and clustering (in which the task can be modeled as assigning values whose only important properties are equalities and

inequalities), the ordinal scale for ranking (in which the order of the values is the important property), and the interval and ratio scales for quantitation (in which differences and ratios are important, respectively).

But we need to add a notion of closeness, that measurement theory does not include. Indeed, we will define two different kinds of closeness: this will allow us to take into account all four abstract tasks, and the related metrics, by just varying two parameters (measurement scale type and kind of closeness). The need for two different kinds of closeness will be detailed in the following, but can be immediately understood at an intuitive level by observing that for both categorization and clustering the scale type is nominal, but the two are clearly different: in the former the *values* used in the assignments are important (otherwise a misclassification occurs), whereas in the latter the equality and inequality relations are important, so any measurement which is *equivalent* for the nominal scale type (another notion provided by measurement theory) is adequate.

So, to summarize, we have three aims in this paper. First, to provide a general definition of metric valid for different abstract tasks such as classification, clustering, ranking, or quantitation (i.e., value prediction). This definition is based on measurement theory, but measurement theory is not enough and we also need to capture the concept of closeness, or proximity; we define two kinds of closeness. Second, to make explicit a correspondence between the four information access abstract tasks and a two dimensional space defined by: (i) the family of the metric (defined on the basis of two different approaches to closeness, based on values and equivalence), and (ii) the scale type to be used to analyse the correspondence between system and gold. Third, to state some theorems showing how our general definitions and axioms specialise into the metric properties defined in the literature for each particular information access task. Thus, the theoretical limitations of specific metrics that have been derived from the basic axioms in the literature can also be derived in our framework.

This paper is structured as follows. In Sect. 2 we extensively survey the previous attempts to formalize metrics properties across various abstract tasks. We then turn to defining our framework, which is based on measurement theory. The reader can find in Appendix A a basic background in measurement theory, a well settled discipline that provides the foundations and tools for our formalization. We include basic definitions and some examples. We ground on measurement theory to generalize the notion of evaluation metric in terms of closeness at specific scale types, as discussed in Sect. 3, where two kinds of closeness are defined. Section 4 focuses on effectiveness metrics: we distinguish two families of metrics; for each of the two we provide a formal definition and state some properties as axioms. In Sects. 5 to 7 we exploit the framework: we state some theorems, that formally capture both general metric properties already proposed in the literature and properties of specific metrics, and that can be derived as particular cases from the definitions and axioms presented in our framework (proofs are in Appendix B). We finally show the generality of our framework by applying it to novel metrics and tasks in Sect. 8. Section 9 summarizes the main results, discusses consequences, assumptions and limits of this study, and sketches future work.

2 Related work

Several authors have proposed formal properties of metrics by defining axiomatics focused on a specific task. Terminology needs some clarification. Different authors have used “properties”, “constraints”, or “axioms”, and there is even some debate on which term is correct. In this paper we privilege the last term, although sometimes we also use

as synonym one of previous two, especially when describing previous work (that we do usually by using the same terminology as in the original). As already mentioned, we distinguish between the concepts of information access *task* and *abstract task*. The former is related with the user context and goals. Examples of tasks could be: searching web pages for generating a report about a topic, recommending products for online sales, spam filtering, sentiment analysis over tweets for reputation analysis, novelty detection in news, etc. All these tasks share the common characteristic that they consists of organising information items (web pages, mails, etc.). In this paper we refer to information items as *documents*.¹ An *abstract task* can be seen as an attempt of formalising the tasks and/or their basic components, and it is related with the characteristics of system outputs and goldstandards (i.e., human judgments/assessments, also called simply *golds*). We focus on the four abstract tasks listed above (Categorization, Clustering, Ranking, and Quantitation), and we survey the main approaches, grouped by abstract task. We also note some details that are useful in the following of the paper. We call *basic axioms* those common to different authors and somehow related to the abstract task but independent from the task. Tables 1, 2, and 3 list the main properties and can be a useful reference; some of the notes in the tables are described in Sect. 6.

2.1 Classification axiomatics

Some authors group classification metrics according to their properties. For instance, Ferri et al. (2009) discriminated between *probabilistic* measures (which consider the deviation from the true probability of errors) and measures based on a *qualitative* understanding of errors (which focus on the idea of utility). The authors do not offer a formal distinction between probabilistic and qualitative measures.

More recently, Sebastiani (2015) proposed eight axioms. The first one is the *Strict Monotonicity* axiom: it states that, given two classification outputs such that they only differ on one decision, i.e., the category of a document, then if one of them is correct on that document, it must be reflected in an increase in its metric score. This idea is also captured by Sokolova's (2006) properties (see below). Sebastiani proved that the traditional F-measure (based on Precision and Recall) does not satisfy this property, as it fails when components of the contingency matrix have zero value. A similar problem is identified for the metric Lam% (Qi et al. 2010). Note, however, that zero values in components of the contingency matrix represent in general a very particular situation. We provide a slight generalization of MON into a Generalized Strict Monotonicity Axiom (GMON) in the following sections.

Sebastiani's second axiom, *Continuous Differentiability*, states that the evaluation measure must be continuous and differentiable over the true positives and true negatives. According to the author, measures fail to satisfy this axiom, again, in the case of zero values in the contingency matrix. Something similar happens with his third and fourth axioms, *Strong Definiteness* and *Weak Definiteness*, which state that the measures must be definable under any gold or system output. One might argue that the *Strict Monotonicity* axiom subsumes these two axioms, because the metric score of every system output must be definable in order to produce a score increase in the *Strict Monotonicity* conditions.

¹ To help intuition, we prefer to use "documents" instead of "items", but let us remark that it is just a matter of terminology and all the results presented in the following are general and hold for any kind of item.

Table 1 The main classification axioms proposed in the literature, with some notes (discussed in more detail in the text) and their correspondence with the axioms and theorems in our framework (presented in the following of the paper)

Classification Axioms	Notes
Sebastiani (2015)	
1. <i>Strict Monotonicity (MON)</i> (*)	Subsumed by GMON
2. Continuous Differentiability (CON)	
3. <i>Strong Definiteness (SDE)</i> (*)	Subsumed by MON
4. <i>Weak Definiteness (WDE)</i> (*)	
5. Fixed Range (FIX)	Non related with quality
6. Robustness to Chance (CHA)	
7. Robustness to Imbalance (IMB)	Task-dependent
8. Symmetry (SYM)	
Sokolova (2006)	
1. Invariance to TP-TN swap	Correspondence with SYM
2. Invariance to TN changes	Complementary to CHA and IMB. Task-dependent
3. <i>Invariance to FP changes</i> (*)	Subsumed by MON
4. <i>Invariance to scaling</i> (*)	Non compatible with MON
<i>Generalized Monotonicity (GMON)</i> (*)	Captured by VOM _N & Theorem 1

Basic axioms are labeled with (*) and in italic

The fifth axiom (*Fixed Range*) sets a restriction about the measure value range. The sixth and seventh axioms (*Robustness to Chance* and *Robustness to Imbalance*) are related to the idea of probabilistic measures proposed by Ferri et al. (2009), and state that random or trivial classifiers must achieve the same score regardless the goldstandard. Measures such as Accuracy, Utility or F-measure (Precision and Recall), although widely adopted, do not satisfy this property. The reason is that actually, there are situations and user contexts in which not every trivial or random classifier has the same effectiveness. For instance, putting randomly just a few mails in the spam directory is less problematic for the user than putting most of mails. The F-measure tackles this aspect by returning a fixed precision for any random output while increasing recall when returning most of e-mails to the user. The eighth and last axiom, *Symmetry*, “enforces the notion that the evaluation measure should be invariant with respect to switching the roles of the class and its complement”. Thus, replacing the positive samples by negative ones in both the system output and the gold produces the same classification score. However, again, this axiom is task-dependent and, therefore, not every metric is designed to satisfy it. For instance, class oriented metrics, such as the F-measure combination of Precision and Recall, do not satisfy it. Utility metrics assign a utility weight to each class, so they do not satisfy it. A system correctly labeling as spam 8 out of 10 spam messages would be useful to the user, even if not perfect, but a system correctly labeling as non-spam 8 out of 10 non-spam messages would probably be unacceptable. For this reason, non symmetric metrics such as class oriented metrics are employed in these tasks.

In an earlier paper, Sokolova (2006) proposed a formal categorisation according to a set of properties based on the invariance of measures under a change in the contingency matrix (TP, TN, FP, FN, i.e., True Positive, True Negative, False Positive, and False Negative, respectively).

Table 2 The main clustering axioms (same notation as in Table 1)

Clustering Axioms	Notes
Dom (2001)	Extended by Rosenberg and Hirschberg's axiomatics
Rosenberg and Hirschberg (2007) (*)	Subsumed by Homogeneity and Completeness
Meila (2003)	
Properties 1, 2, 3, 5, 6, 8, 9, 12	Non related with quality aspects
Properties 4 and 7	Related with Cluster Size versus Quantity
Properties 10 and 11 (*)	Related with Completeness
Amigó et al. (2009)	
1. Cluster homogeneity (*)	Generalized in GHC
2. Cluster completeness (*)	
3. Rag bag	Task-dependent
4. Cluster size versus quantity	
Generalized Homogeneity / Completeness (GHC) (*)	Captured by EOM _N & Theorem 2

These properties have a correspondence with axioms proposed by other authors. The invariance under the swap of TP with TN and FP with FN corresponds to Sebastiani's Symmetry axiom. The second property is the invariance under the change in TN when all other matrix entries remain the same. This property characterises the class-oriented measures (Precision and Recall). If the measure is not sensitive to one of the components, then increasing the amount of returned documents when the document classification is random, can be always beneficial. This property is complementary to the sixth and seventh Sebastiani's axioms; thus, this property is also task-dependent. The next property is the invariance under the change in FP when all other matrix entries remain the same. The non-invariance is necessary if the measure satisfies the Strict Monotonicity axiom. The fourth and last property is the invariance under the classification scaling. According to the author, it is only satisfied by Precision, which is a partial measure that does not satisfy the Strict Monotonicity axiom.

In summary, we see that some axioms and properties are related, equivalent, or subsumed, and others are task-dependent. The only exception we find is the Strict Monotonicity axiom, which is common across all authors and is generally satisfied by most metrics. According to this analysis, we consider Strict Monotonicity as the unique commonly accepted basic axiom for the classification abstract task.

2.2 Clustering axiomatics

Dom (2001) proposed five formal desirable properties for clustering metrics. These were later extended to seven by Rosenberg and Hirschberg (2007). Basically, this axiomatics consists of stating a bijective correspondence between clusters and classes in the gold. It assumes a set of useful clusters with high correspondence with classes (peer to peer) and

Table 3 The main ranking axioms (same notation as in Table 1)

Ranking Axioms	Notes
Moffat (2013)	
1. Boundedness	Non related with quality aspects
2. Monotonicity	Task-dependent
3. <i>Convergence</i> (*)	Non compatible with Confidence axiom
4. Top-weightedness	Subsumed by Priority axiom
5. Localization	Task-dependent
6. Completeness	Non general. Only for deepness threshold based metrics
7. Realizability	About definability
Ferrante et al. (2015)	
1. <i>Replacement</i> (*)	Non related with ranking quality
2. <i>Swapping</i> (*)	Subsumed by the Priority axiom
Amigó et al. (2013)	
1. <i>Priority (PRI)</i> (*)	Captured by EOM_0 & Theorem 3
2. Deepness	Task-dependent
3. Closeness Threshold	
4. Deepness Threshold	
5. Confidence	

a set of noisy (small) clusters. The axioms state that increasing the amount of noisy clusters, splitting or joining useful clusters decrease the score. These properties are implicitly subsumed by the Generalized Homogeneity/Completeness axiom that we introduce below, given that these movements require breaking correct relationships and increasing the amount of incorrect relationships.

Meila (2003) proposed an entropy-based metric (Variation Information) and listed twelve desirable properties associated with it. Most of these properties are not directly related to the quality aspects captured by a metric, but rather to other intrinsic features such as the ability to scale or computational cost (Amigó et al. 2009). The exceptions are the properties 4 and 7, related with *Cluster size versus Quantity* (task-dependent), and properties 10 and 11, related with *Completeness*.

Amigó et al. (2009) proposed an axiomatics consisting of four constraints that, by focusing on extreme situations in which one system output should outperform another, capture the essence of previously proposed axiomatics (Dom 2001; Meila 2003):

- *Cluster Homogeneity*: given a certain system output document distribution, splitting documents that do not belong to the same class must increase the output quality. This restriction was first proposed by Rosenberg and Hirschberg (2007). Although it seems a very basic constraint, measures based on editing distance do not satisfy it (Amigó et al. 2009).
- *Cluster Completeness*: the counterpart to the first constraint is that documents belonging to the same class should be in the same cluster. This intuition is also captured by Dom's constraints. Measures based on set matching, such as Purity and Inverse Purity, do not satisfy this constraint.

- *Rag Bag*: introducing disorder into a disordered cluster (rag bag) is less harmful than into a clean cluster. In general, all traditional measures fail to comply with this constraint. However, it can be considered task-dependent, for example, in an early alarm detection task, considering a few related messages in isolation could be crucial.
- *Cluster size versus quantity*: a small error in a big cluster is preferable to a large number of small errors in small clusters. This constraint prevents the problem that measures based on counting pairs (Amigó et al. 2009; Meila 2003; Halkidi et al. 2001) overweight big clusters (these measures are sensitive to the combinatorial explosion of pairs in big clusters, and fail on this constraint). Although this principle is shared by several authors (Amigó et al. 2009; Dom 2001; Meila 2003), we can consider a task in which this axiom is not mandatory, as one could be interested in penalizing errors in large clusters more than multiple errors in small clusters.

Cluster Homogeneity and Cluster Completeness constraints can be generalized into a single one: given two identical clustering outputs with the exception that the second output contains correct clustering relationships that do not appear in the first output, or it does not contain incorrect clustering relationships that appear in the first system, then the metric must strictly increase. In the following sections we formalise this as *Generalized Homogeneity/Completeness (GHC)*.

In summary, we can conclude that the basic axioms which are shared by different analyses are *Completeness* and *Homogeneity* which can be generalized into a unique axiom.

2.3 Ranking axiomatics

Most of the work on ranking axiomatics has been developed in the context of IR (as discussed in the first part of this section). There has been some discussion about ordinal classification too (as discussed in the last part). It is however important to understand that, by being focused on those concrete tasks, researchers have proposed, if not taken for granted, some properties, that do not apply *for the abstract task of ranking*. For example, it might sound surprising but top-heaviness is not mandatory for the abstract task of ranking. It is possible to imagine concrete ranking tasks that do not reward correctness in earlier rank positions more than in later ones: to provide a simple example, if one has to evaluate an approximate algorithm alphabetically sorting an array, Kendall's Tau correlation would be a reasonable measure, although it is not top-heavy — and indeed it would not make any sense to reward the sorting algorithms that are more correct for “A” than for “Z”.

In one of the early works on formalizing IR evaluation, Van Rijsbergen (1974) already suggested to use measurement theory (as we do extensively in this paper). However, that seminal work does not state formal properties for simple evaluation measures, but for the combination of them (e.g., the well known F-measure), using the Conjoint Measurement Theory. Other authors tried to exploit measurement theory and to define a notion of similarity between measurements when formalizing IR (Busin and Mizzaro 2013; Maddalena and Mizzaro 2014; Ferrante et al. 2015). This paper extends that idea, but using measurement theory to state definitions and axioms for evaluation measures equally valid for different information access abstract tasks, not just IR.

Moffat (2013) listed seven properties that IR metrics should satisfy. The first one is *Boundedness*. It is not about quality; it requires the existence of a bounded range of scores. The second one is called *Monotonicity* and states that if a ranking of length k is extended so that $k + 1$ elements are included, the metric value never decreases. This second property

is task-dependent: in some situations, reducing the size of the returned document list could be useful, for instance if it contains only irrelevant documents. In addition, according to the author, this property gets in contradiction with the next one, the *Convergence* property, that reflects the basic principle that relevant documents must occur above irrelevant documents in the system output ranking. The property states that swapping two documents in the ranking in concordance with the relevance judgements strictly increases the metric scores. The fourth property, *Top-weightedness*, explicits that in IR the first positions in the rank are the most important ones. As anticipated at the beginning of this section, this constraint, although widely accepted, is task-dependent, because it is related with the cost of exploring the ranking produced by the system, i.e., the probability that the user actually explores each ranking position. Usually it is reasonable to assume that this probability decreases by going down the rank, but one might imagine situations where this is not true and a very persistent user explores all the ranking positions. The fifth one is referred as *Localisation*: a metric value at a given rank position k should depend only on the documents in the first k positions. This axiom can be applied only to metrics that assume a certain deepness threshold as input parameter. That is, it is not a general axiom. The sixth property, *Completeness*, states that the metric must be definable even when there are no relevant documents in the collection. Finally, *Realisability* states that the maximal score can be achieved even when there exists only one relevant document in the collection. This property is related with the normalisation of scores across test cases.

Ferrante et al. (2015) proposed two axioms: *Replacement* (replacing an irrelevant document in the output ranking with a relevant one increases the score) and *Swapping* (swapping two documents in the ranking output in concordance with the gold annotation relative relevance increases the score). In fact, if we assume that documents out of the output ranking are located together in an additional last ranking position, then Replacement is subsumed by Swapping. In addition, Swapping is equivalent to *Convergence*.

Amigó et al. (2013) proposed five axioms. The first one (*Priority (PRI)*) states again that swapping documents in a correct way increases the score. It is similar to Ferrante's Swapping axiom, and equivalent to it, although somehow more relaxed since it requires a metric score increase only when the ranking position of the swapped documents are contiguous. The next three axioms are related with assumptions about the user task: deeper positions in the ranking are less likely to be explored (*Deepness* axiom); there is an area at the top of the ranking which is always explored by the user (*Closeness Threshold* axiom); and there is an area deep enough that is never explored by the user (*Deepness Threshold* axiom). The last two axioms are formalised in terms of comparing one relevant document in the first position, n relevant documents in the $2n$ first ranking positions, and a huge amount of relevant documents after a huge amount of irrelevant documents. The fifth axiom states that given a ranking containing only irrelevant items, the shorter the ranking the higher the metric score (*Confidence* axiom). This axiom is also task-dependent. We could think that reducing the ranking length, instead of avoiding user effort, adds uncertainty about the possibility of finding relevant documents at lower ranks.

In summary, according to our analysis, most of the axioms and properties are task-dependent. The basic axioms, common to all studies, are related with the correctness of priority relationships: the Swapping constraint proposed by both Ferrante et al. (2015) and Moffat (2013), and its relaxed version, the Priority axiom proposed by Amigó et al. (2013).

We also briefly mention the task of assigning items to categories which have an order, named *Ordinal Classification*: is a sort of mixture between classification and ranking. It is a quite popular task in some situations: besides the common assignment of "stars" in several Web

reviews sites, let us consider the polarity detection task, in which text fragments must be classified in terms of sentiment analysis according to a few categories such as “*Very positive*”, “*Positive*”, “*Neutral*”, “*Negative*” and “*Very negative*”. The task is defined in an ordinal manner, but the perfect system should return exactly the same values, so it is not a ranking problem like traditional IR. Then, some evaluation campaigns have applied a classification oriented metric, such as in Barbieri et al. (2016). In other campaigns, systems were evaluated with ranking metrics, such as in Amigó et al. (2013); in other ones, different tasks were defined in concordance with different metrics (ranking or classification), such as in Rosenthal et al. (2017).

We can see a similar situation in semantic textual similarity (Agirre et al. 2015) in which text pairs must be categorised according to a few classes (high / average similarity, etc.): the organizers used Pearson correlation. The prediction of stars in product recommendation also fits into this case: sorting products in a correct way is desirable, as well as assigning the correct amount of stars. In fact, recommender systems were initially evaluated in terms of accuracy, before the community began to work under ranking based metrics. We can find a few studies in the literature that analysed the behavior of some traditional metrics in this task such as Mean Average Error, Mean Squared Error, linear correlation, or Accuracy with n (Gaudette and Japkowicz 2009), proposed a method to make Ordinal Classification metrics robust to imbalance (Baccianella et al. 2009), and analysed the suitability of traditional metrics by means of a particular case and proposed the Ordinal Classification Index metric (Cardoso and Sousa 2011). However, differently from the previous abstract tasks, these studies do not define properties to be satisfied, and in general there is not a clear choice of the metric to be used when predicting a few labels that keep an ordinal relationship.

2.4 Quantitation axiomatics

Quantitation,² i.e., the abstract task of assigning numeric values to documents, is perhaps less common, but it is not only theoretical and some examples can be found. In the Semantic Textual Similarity task at SemEval-2016 (Agirre et al. 2016), systems are asked to return a numeric value predicting the similarity between two snippet of texts. In some sentiment polarity detection tasks, the absolute polarity values returned by systems are compared with the reference values. In both cases the stated goal consists of maximising the linear correlation between system outputs and golds. The most frequent metric in these cases is the Pearson coefficient, although this choice is being criticised (Reimers et al. 2016) and might change in the future.

In other cases the goal consists of predicting the exact values. One example is the proposal to evaluate information retrieval systems not only on the basis of the rank of the retrieved documents, but on the basis of the numeric relevance values assigned to document. With this approach, and assuming a continuous notion of relevance, Della Mea and Mizzaro (2004) proposed ADM (Average Distance Measure). More recently, magnitude estimation has been proposed as a technique to gather relevance assessment on a ratio scale (Maddalena et al. 2017). One might even claim that there is a trend to go beyond the classical (category relevance and ranking retrieval) situation, although we are not aware of any attempt to capture the properties of metrics for quantitation. Therefore we include this abstract task in our analysis.

² To avoid ambiguity, we use “quantitation” instead of its synonym “quantification” [https://en.wikipedia.org/wiki/Quantification_\(science\)](https://en.wikipedia.org/wiki/Quantification_(science)), that in data and text mining “consists in providing an aggregate estimation (e.g., the class distribution in a classification problem) for unseen test sets, applying a model that is trained using a training set with a different data distribution” (González et al. 2017, p. 74).

3 Measurement theory and closeness

Measurement theory is briefly recalled in Sects. 3.1 and 3.2 (for more details see Appendix A), where we also discuss the notion of closeness. We then outline the structure of our framework in Sect. 3.3. Sections 3.4–3.6 provide some definitions, and Sect. 3.7 provides an example (and the reader might find it useful to go back to Sects. 3.4–3.6 after having read it).

3.1 Measurement theory

Appendix A recalls some of the basic concepts of measurement theory, that we assume as known in the following: assignment, measurement, scale types and permissible transformation functions, equivalence, and meaningfulness (see Definitions A.1–A.4 and A.6). The reader familiar with measurement theory can probably just skim the appendix to get acquainted with our notation, or maybe even skip it and refer to specific parts when needed.

Briefly, measurement theory studies the properties of value assignments to objects like, for instance, temperature, height, and distance. At the core of measurement theory there is the notion of scale type. The classical scale types are nominal, ordinal, interval, and ratio. At each scale type, some relationships between values make sense and others are not meaningful. For instance, at the nominal scale type only equality and inequality of values can be taken into account, whereas considering the ratio or the interval between values makes no sense (e.g., “red” divided by “green”). For each scale type, a set of permissible transformation functions is defined. These, when applied to a measurement, determine which are the measurements that, though different, are equivalent, i.e., carry the same meaning. For example, for the ordinal scale type, value assignments that order objects in the same way are equivalent, and the set of permissible transformation functions are the strictly monotonically increasing functions.

This matches with our abstract tasks: in general, we can say that systems assign values to items (relevance, topics, categories, priority, etc.), and evaluation consists in comparing the system output assignment against the human annotated assignment (gold). As an example, suppose that we want to categorize some documents into “physics”, “biology”, and “social sciences”. This is a classification problem, and there are no ordinal or interval relationships between classes. The goal consists in maximizing the amount of equalities between system output values and gold values, and there is no consideration about the range or interval of errors. In other words, the predicted categories are either accurate or not. In other terms, classification problems can be mapped onto the nominal scale type. Similarly, ranking problems can be mapped onto the ordinal scale type, and quantitation problems onto the interval and ratio scales. However the situation is slightly more complex; this can be seen when considering that clustering can be mapped onto the nominal scale type as well, but it is an intrinsically different problem from classification. Anticipating what we will discuss in detail in the following, we model this difference on the basis of the kind of closeness that can be defined between two measurements: for classification we seek for equal measurements; for clustering we seek for equivalent measurements. We will also discuss that it is not even necessary to assume that system outputs and golds are measurements; assignment is enough.

Incidentally, we remark that by exploiting the basic concepts of measurement theory, like scale types and meaningful statements, it would be possible to directly derive some

consequences on effectiveness metrics. For example, the well known Mean Reciprocal Rank (MRR) metric is computed by considering the rank of the first relevant document, computing its reciprocal, and averaging these values. But by doing so, one is neither applying permissible transformation functions, nor deriving meaningful statements, for the scale type at hand, which is the ordinal one; a similar remark has been made by Fuhr (2018). Also the widely adopted Normalized Discounted Cumulative Gain (nDCG) metric transforms an ordinal relevance scale (e.g., Highly relevant, Relevant, Marginally relevant, Not relevant) into numerical gains which are on a ratio scale type, and this is intrinsically arbitrary.

But we believe that the consequences of measurement theory on evaluation are deeper; in the rest of the paper we discuss those more foundational aspects.

3.2 Closeness

Measurement theory focusses on equivalence rather than closeness. For instance, according to Suppes and Zinnes (1963) the two first fundamental problems in measurement theory are the *Representation Theorem* and *Uniqueness*. Both are related with finding homomorphisms between the empirical and numerical structures, thus studying which assignments are indeed measurements. The third problem is the *Meaningfulness Problem* (see Sect. A.4). There is nothing that discusses the similarity, or distance, or closeness of two assignments or measurements. However, the main goal of evaluation is to compare system outputs against gold standards or references, which are almost never equivalent. In fact, equivalence would mean perfect effectiveness, which is an extremely rare event in information access. So, a simple binary comparison (equivalent versus non equivalent) is not enough, and, in this sense, the concept of *closeness between two assignments* (and, consequently, measurements) needs to be added to measurement theory concepts to formalise the evaluation scenario. As we will see in Sect. 4, this will allow to define the notion of evaluation metrics. Going back to the classification problem of the example in Sect. 3.1, this means that we need to model the closeness between system output values and gold values at the nominal scale type.

As another example, suppose that we are interested in grouping documents correctly. That is, the system must assign to documents about physics the same value, but different from biology documents. In this case, the document tags generated by the system (assigned values) are not necessarily equal to the category identifiers provided by humans in the gold. For this, we can exploit the notion of *equivalence* in measurement theory. That is, two assignments are equivalent at the nominal scale type if they keep the same equality relationships between objects. For instance, the assignments ($A = 1, B = 1, C = 2$) and ($A = 3, B = 3, C = 1$) are equivalent at the nominal scale type, given that, in both cases, $A = B \neq C$ is true. This matches with our clustering problem. In terms of measurement theory we can say that the goal of clustering evaluation is to quantify how close is the system output to be equivalent to the gold at the nominal scale type.

Let us finally consider the abstract task of sorting documents according to their relevance, the classical document ranking problem. In this case, the evaluation must consider the ordinal relationships between assigned values, while the interval or ratio between assigned values is not important. That is, documents must be sorted in the same way as in the gold, i.e., relevant documents earlier than irrelevant documents. In terms of measurement theory, the goal of ranking is generating a relevance assignment equivalent to the gold. We can interpret document ranking as an ordinal equivalence oriented problem.

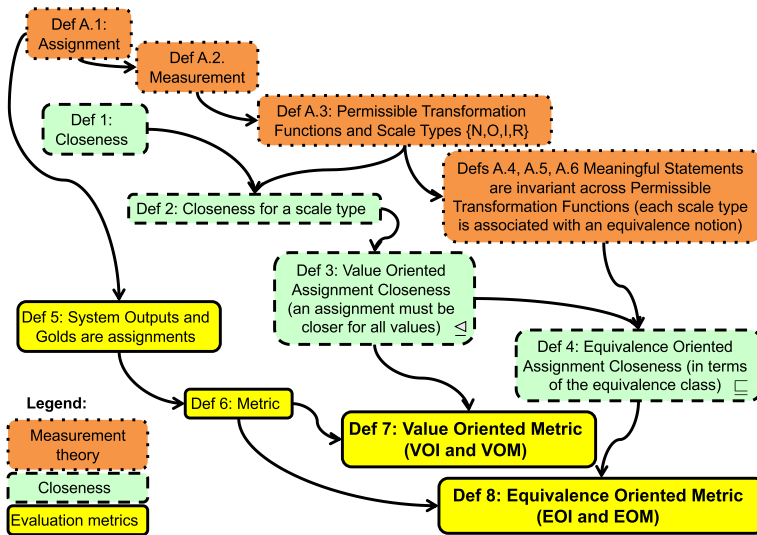


Fig. 1 Overall structure of the framework: from basic definitions of measurement theory, through closeness, to the two families of metrics

We state two general definitions of evaluation metric, value-oriented and equivalence-oriented, and we prove that, depending on the scale type, existing metrics and their desirable properties defined in the literature for each abstract task fit into the corresponding definition. But then, more situations appear spontaneously in the proposed model. For instance, in polarity detection, (positive, neutral, negative) categories present an ordinal relationship, while the intervals between polarity levels are undefined. However, unlike in ranking problems, predicting the specific polarity category of documents must be rewarded. This is an Ordinal Classification problem (see Sect. 2.3), but there is not a consensus about how this problem must be formalized. Our framework identifies it as a value-oriented ordinal problem, fitting in our definition.

The proposed model goes further, and includes abstract tasks at interval and ratio scales. The model also allows us to analyze evaluation metrics, and to understand when they can be used appropriately. For example, we prove in this paper that Pearson coefficient fits into our equivalence-oriented metric definition at the interval scale type, the popular cosine distance fits into the equivalence-oriented metric definition at the ratio scale type, and the commonly used error rate or mean average error fit into the value-oriented metric definition at the ratio or interval scale type.

3.3 Structure of the framework

By exploiting some basic results of measurement theory, as well as the notion of closeness, we now turn to defining our framework. Figure 1 sketches the overall structure of the framework and can be a useful reference in the following. We ground on some definitions from measurement theory (orange boxes with dotted borders); we have briefly recalled them above but the reader is referred to the corresponding Definitions A.1–A.6 in Appendix A. Then, we analyse the notion of proximity, or *closeness* (dashed green boxes). We start from

a simple but general definition of closeness (Definition 1) and we particularize it for the four classical scale types (Definition 2). On this basis, we define two kinds of closeness measures for assignments, depending on whether they focus on value matching (Definition 3) or on measurement equivalence (Definition 4). These definitions can be particularised for any scale type: nominal, ordinal, interval, or ratio. Then, moving to the yellow boxes, we define system outputs and goods as assignments of numeric values to items (Definition 5); let us remark that this is done for compatibility with measurement theory and without loss of generality as long as we consider the abstract tasks. Finally, the definition of metric in general is directly derived from the notion of closeness (Definition 6), as well as the the two specific definitions of value- and equivalence-oriented metrics (Definitions 7 and 8).

3.4 Defining closeness for a scale type

In this subsection, we start from a simple but general definition of closeness between single values, and then we propose a definition dependent on the scale types. Exploiting these definitions, a more intuitive description of closeness for the scale types is then derived in Lemma 1; this will be used to define closeness between assignments in the next subsections.

A first important remark is that it is possible to define different distance functions, and therefore, multiple closeness notions. There is some unavoidable arbitrariness. A second remark is that the closeness between values also depends on the scale type of reference. For the nominal scale type, two values are similar when they are equal (e.g., since $3 \neq 4 = 4$, 4 is more similar to 4 than 3). For the ordinal scale type, we observe relative closeness when values are in sequence (e.g., since $3 < 4 < 5$, 3 is more similar to 4 than to 5). For the interval scale type, when the absolute difference is lower (e.g., since $|3 - 5| < |1000 - 5|$, 3 is more similar to 5 than 1000). Notice that due to the subsumption effect across scale types (Formulas (30) and (31)), each assertion is valid for the other higher scale types (e.g., if $3 \neq 4 = 4$ then $3 < 4 \leq 4$ and $|3 - 4| > |4 - 4|$).

For our framework a simple definition of closeness is sufficient, based on a distance between two values $x, y \in \mathbb{R}$ computed as $|x - y|$. Moreover, we are interested in comparing closeness values, not in precise closeness values. Therefore we do not define a function returning a closeness value, but simply the following relationship.

Definition 1 (*Closeness*) Let $x, y, r \in \mathbb{R}$. We say that x is (strictly) closer to a reference value r than y , and we write $x \leq^r y$ ($x <^r y$ for the strict case), if and only if:³

$$x \leq^r y \iff |r - x| \leq |r - y| \tag{1}$$

$$x <^r y \iff x \leq^r y \wedge \neg(y \leq^r x) \iff |r - x| < |r - y|. \tag{2}$$

We now particularize closeness for a certain scale type. We refer to the four classic scale types, i.e., *Nominal* (denoted with N from now on), *Ordinal* (O), *Interval* (I), and *Ratio* (R). There is a natural order on the scale types, going from the lowest scale type N to

³ It is easy to see that the last equivalence in (2) derives simply by applying (1) and observing that

$$\begin{aligned} x \leq^r y \wedge \neg(y \leq^r x) &\iff |r - x| \leq |r - y| \wedge \neg|r - y| \leq |r - x| \\ &\iff |r - x| \leq |r - y| \wedge |r - x| < |r - y| \iff |r - x| < |r - y| \iff x <^r y. \end{aligned}$$

Table 4 Examples for closeness and strict closeness at different scale types

	-2	-1	-0.5	0	0.5	1	2
$v \stackrel{\mathcal{N}}{\leq} 1$	○	○	○	○	○	○	○
$v \stackrel{\mathcal{O}}{\leq} 1$	○	○	○	○	○	○	×
$v \stackrel{\mathcal{I,R}}{\leq} 1$	×	○	○	○	○	○	×
$v \stackrel{\mathcal{N}}{<} 1$	×	×	×	○	×	×	×
$v \stackrel{\mathcal{O}}{<} 1$	×	×	×	○	○	×	×
$v \stackrel{\mathcal{I,R}}{<} 1$	×	×	○	○	○	×	×

the highest scale type R. This order is derived from the inclusion chain of the permissible transformation functions of the four scale types: if \mathcal{F}_T denotes the set of permissible transformation functions for the scale type T, then $\mathcal{F}_R \subset \mathcal{F}_I \subset \mathcal{F}_O \subset \mathcal{F}_N$. This allows us to write $N < O < I < R$ and to speak of higher and lower scale types accordingly. See Appendix A for further details.

Definition 2 (*Closeness for a scale type*) Let $x, y, r \in \mathbb{R}$, and \mathcal{F}_T the set of permissible transformation functions for the scale type T. We say that x is closer to a reference r than y for a certain scale type T, and we write $x \leq_T^r y$ if and only if it is closer for at least one permissible transformation function in \mathcal{F}_T :

$$x \leq_T^r y \iff \exists f \in \mathcal{F}_T (f(x) \leq^{f(r)} f(y)). \tag{3}$$

The associated strict relationship is:

$$x <_T^r y \iff x \leq_T^r y \wedge \neg(y \leq_T^r x). \tag{4}$$

Given a fixed reference r , non-strict closeness is a binary relationship that satisfies reflexivity ($\forall x (x \leq_T^r x)$), transitivity ($\forall x, y, z (x \leq_T^r y \wedge y \leq_T^r z \implies x \leq_T^r z)$), and is a connex relation ($\forall x, y (x \leq_T^r y \vee y \leq_T^r x)$). Non-strict closeness is not a total order since it is not anti-symmetric, as two different values can be equally non-strict close to the reference. The strict closeness relationship is irreflexive, transitive, asymmetric ($\forall x, y (x <_T^r y \implies \neg(y <_T^r x))$), and acyclic.

Table 4 illustrates the behavior of these definition by analyzing whether the value v is non-strictly or strictly closer to the reference 0 than 1 for any scale type T (i.e., $v \leq_T^0 1$ and $v <_T^0 1$). Each column is associated with a value v , and each row with non-strict or strict closeness under different scale types. In other terms, we ask whether a given value v is closer to 0 than 1 for a given scale type; the upper and lower parts of the table discuss strict and non-strict closeness, respectively. For instance, looking at the first row, all values are non-strictly closer (actually, equally closer) to 0 than 1 for the nominal scale type, since a permissible transformation function, i.e., a bijective function, can be found that transforms any value into any other one, including 0; looking at the fourth row, only the zero value is strictly closer to 0 than 1 for the nominal scale type. For the ordinal scale type (second row), only 2 can not be non-strictly closer to 0 than 1 (because 2 is farther away to 0 than 1 for any monotonic transformation); looking at the fifth row, any value in $[0, 1)$ is strictly closer. For the interval and ratio scale types, any value in $(-1, 1)$ is strictly closer to 0 than 1. In general, the table reflects the subsumption of permissible transformation functions across scale types (Formulas (30) and (31)). The effect is that, when considering higher

Table 5 The conditions for closeness and strict closeness in Lemma 1

T	$x \leq_T^c y$	$x \leq_T^s y$
N	$(r \neq y \vee r = x)$	$(r = x \wedge r \neq y)$
0	$\neg((r \geq y > x) \vee (x > y \geq r))$	$(r \geq x > y) \vee (y > x \geq r)$
I, R	$ r - x \leq r - y $	$ r - x < r - y $

scale types, the number of strict closeness relationships increases, whereas the number of non-strict closeness relationships decreases.

This definition of (strict) closeness is general and valid for any scale type. We can instantiate it into the four scale types as shown by the following lemma; this form helps intuition and it will be useful to develop the formal proofs in the following (all the proofs are in Appendix B).

Lemma 1 (Closeness for the four scale types) *Let $x, y, r \in \mathbb{R}$. The value x is (strictly) closer to a reference r than y for each scale type T respectively if and only if the conditions in Table 5 are satisfied.*

We now turn to closeness for assignments. We define two kinds of closeness (value-oriented and equivalence-oriented), whose meaning is discussed in the example in Sect. 3.7.

3.5 Value-oriented assignment closeness

On the the basis of value closeness, we define a first closeness notion between assignments for a certain scale type. Consistently with Definition A.1, we denote assignments with $\omega, \omega', \omega_i$, etc.

Definition 3 (Value-oriented assignment closeness) *Given a set of objects \mathcal{D} , an assignment ω is value-closer to a reference assignment ρ than another assignment ω' for the scale type T (we write $\omega \leq_T^\rho \omega'$ and we speak of Value-Oriented Assignment Closeness) if and only if for every value (i.e., objects in \mathcal{D}) ω is closer to ρ than ω' for the scale type T :*

$$\omega \leq_T^\rho \omega' \iff \forall d \in \mathcal{D} \left(\omega(d) \leq_T^{\rho(d)} \omega'(d) \right). \tag{5}$$

Moreover, we say that an assignment ω is strictly value-closer to a reference assignment ρ than another assignment ω' for the scale type T if:

$$\omega \triangleleft_T^\rho \omega' \iff \omega \leq_T^\rho \omega' \wedge \neg(\omega' \leq_T^\rho \omega).$$

Note that strict value-closeness can be expressed as:

$$\begin{aligned} \omega \triangleleft_T^\rho \omega' &\iff \forall d \in \mathcal{D} \left(\omega(d) \leq_T^{\rho(d)} \omega'(d) \right) \wedge \neg \left(\forall d \in \mathcal{D} \left(\omega'(d) \leq_T^{\rho(d)} \omega(d) \right) \right) \\ &\iff \forall d \in \mathcal{D} \left(\omega(d) \leq_T^{\rho(d)} \omega'(d) \right) \wedge \exists d \in \mathcal{D} \left(\neg \left(\omega'(d) \leq_T^{\rho(d)} \omega(d) \right) \right), \end{aligned}$$

i.e., by applying (4) for any scale type $T \in \{N, 0, I, R\}$,

Table 6 The example described in the text

\mathcal{D}	ρ	ω_1	ω_2	ω_3	ω_4
o_1	10	1	11	12	12
o_2	20	2	20	20	22
o_3	30	3	30	30	32

$$\omega \triangleleft_T^\rho \omega' \iff \forall d \in \mathcal{D} \left(\omega(d) \leq_T^{\rho(d)} \omega'(d) \right) \wedge \exists d \in \mathcal{D} \left(\omega(d) <_T^{\rho(d)} \omega'(d) \right). \tag{6}$$

We use “value-closer” (and not simply “closer”) because we now define another closeness notion for assignments, before discussing an example in Sect. 3.7.

3.6 Equivalence-oriented assignment closeness

We formalise the closeness between assignments in terms of their equivalence class, rather than value correspondence.

Definition 4 (*Equivalence-oriented assignment closeness*) Given a set of objects \mathcal{D} , an assignment ω is equivalence-closer to a reference ρ than another assignment ω' for the scale type T (we write $\omega \sqsubseteq_T^\rho \omega'$ and we speak of *Equivalence-Oriented Assignment Closeness*) if and only if for every assignment ω'_i in the equivalence class of ω' , there exists at least one assignment in the equivalence class of ω that is value-closer to ρ for the scale type T :

$$\omega \sqsubseteq_T^\rho \omega' \iff \forall \omega'_i \in [\omega']_T \left(\exists \omega_i \in [\omega]_T \left(\omega_i \triangleleft_T^\rho \omega'_i \right) \right). \tag{7}$$

The strict closeness is analogous to Definition 3:

$$\omega \sqsubset_T^\rho \omega' \iff \omega \sqsubseteq_T^\rho \omega' \wedge \neg(\omega' \sqsubseteq_T^\rho \omega). \tag{8}$$

Two assignments are closer according to Definition 3 if they tend to assign similar values, and according to Definition 4 if there exists a permissible transformation function that makes them assign similar values. In general, Definition 3 is useful for tasks like “measure temperature using the Celsius scale”; Definition 4 for tasks like “measure temperature using a scale of the interval scale type”. Referring to our abstract tasks, as we will discuss in detail in the following, the former is useful for categorization, where the assigned values are important; the latter for clustering, where cluster labels can be changed without affecting the result. The two kinds of assignment closeness lead to two different families of metrics, as discussed in Sect. 4; we first provide an intuitive example.

3.7 An example

As an example, consider the situation in Table 6, representing some assignments of three objects in $\mathcal{D} = \{o_1, o_2, o_3\}$: ρ is the reference assignment, and the ω_i are some different assignments. Now, if the scale type is \mathbb{N} or \mathbb{O} , all the assignments are equivalent. Therefore, these ω_i are all equally close to the reference ρ in terms of equivalence-closeness.

However, the situation changes when looking at value-oriented closeness, since ω_2 and ω_3 are strictly value-closer to ρ than the other assignments, given that they achieve equality for objects o_2 and o_3 . Thus, for the \mathbb{N} scale type:⁴

$$\omega_2, \omega_3 \triangleleft_{\mathbb{N}}^{\rho} \omega_1, \omega_4. \tag{9}$$

If the scale type is \mathbb{O} the situation is slightly more complex. In addition to the previous relationships, now ω_2 is closer to ρ than ω_3 ($\omega_2 \triangleleft_{\mathbb{O}}^{\rho} \omega_3$) since the value 11 is ordinal closer to 10 than 12. At interval (or ratio) scale type, in addition, $\omega_4 \triangleleft_{\mathbb{I}}^{\rho} \omega_1$, and therefore (where \mathbb{T} is \mathbb{I} or \mathbb{R})

$$\omega_2 \triangleleft_{\mathbb{T}}^{\rho} \omega_3 \triangleleft_{\mathbb{T}}^{\rho} \omega_4 \triangleleft_{\mathbb{T}}^{\rho} \omega_1. \tag{10}$$

The meaningful statements of the interval scale type capture differences between assignments that are not captured at ordinal or nominal scale types.

When considering again equivalence-oriented assignment closeness, we obtain a different, and contradictory, outcome: ω_1 and ω_4 are closer to ρ than ω_2 and ω_3 , i.e.,

$$\omega_1, \omega_4 \sqsubset_{\mathbb{I}}^{\rho} \omega_2, \omega_3. \tag{11}$$

To see why, consider that for any linear transformation applied to ω_2 or ω_3 , we can define a transformation for ω_1 or ω_4 to make them closer to ρ : for instance $\omega'_1 = \omega_1 * 10$ and $\omega'_4 = \omega_4 - 2$. Finally, if the scale type is \mathbb{R} , ω_1 is equivalence-closer to ρ than ω_4 , i.e.,

$$\omega_1 \sqsubset_{\mathbb{R}}^{\rho} \omega_4. \tag{12}$$

It is clear that which assignment to prefer depends on the scale type, the abstract task, and whether value-oriented or equivalence-oriented closeness is used.

4 Metrics: two families, eight classes

We can now turn to analyse effectiveness metrics. As it can be seen in the following, the previous framework based on the inclusion of closeness in measurement theory allows general definitions and theorems (in the next sections).

We first define some notation and provide a basic definition of metrics (in Sect. 4.1); on this basis, and exploiting the two kinds of closeness notions, namely value- and equivalence-oriented (Definitions 3 in Sects. 3.5 and 4 in Sect. 3.6, respectively), we distinguish between *two families of metrics*, namely value- and equivalence-oriented metrics (in Sects. 4.2 and 4.3, respectively). We then revisit the example of Sect. 3.7 (in Sect. 4.4) and classify the metrics into eight classes (in Sect. 4.5).

⁴ We are slightly abusing the notation here: the meaning of Formula (9) is that $\omega_2 \triangleleft_{\mathbb{N}}^{\rho} \omega_1$, $\omega_2 \triangleleft_{\mathbb{N}}^{\rho} \omega_4$, $\omega_3 \triangleleft_{\mathbb{N}}^{\rho} \omega_1$, and $\omega_3 \triangleleft_{\mathbb{N}}^{\rho} \omega_4$. A similar remark holds for the following Formula (11).

4.1 System outputs, gold standards, and metrics

The first step is to formalise system outputs and golds as assignments: they can be relevance judgments, assigned categories, etc. We use α for golds and σ for system outputs (α is a mnemonic for assessment, σ for system); thus $\alpha, \sigma \in \Omega$ (see Definition A.1).

Definition 5 (*System output and gold*) A system output σ or a gold α is an assignment from a set of documents \mathcal{D} to real numbers \mathbb{R} :

$$\sigma : \mathcal{D} \longrightarrow \mathbb{R} \text{ and } \alpha : \mathcal{D} \longrightarrow \mathbb{R}.$$

This approach is different from the classical approach by van Rijsbergen (1981) who focussed on considering measuring retrieval effectiveness itself as a measurement, as well as from the recent proposal by Ferrante et al. (2017), who consider evaluation metrics as measurements and set to determine if an IR evaluation metric is a measure on an interval scale. We do not represent an effectiveness metric as a measurement. Our approach is more similar to the already cited work of Busin and Mizzaro (2013) and Maddalena and Mizzaro (2014), but with an important difference. That previous work modeled system outputs and golds as measurements; however, this is not necessary in our approach, where they are simply assignments. This is an important simplification. We also remark that we are not the only ones to represent Golds as assignments. For example, Ferrante et al. (2019, Section 4) write: “the ground-truth GT is a map which assigns a relevance degree $rel \in REL$ to a document d with respect to a topic t ”. We extend that approach by applying it to (i) system outputs and (ii) other abstract tasks beyond IR (ranking).

On the basis of Definition 5 we can represent any system output and gold. For example, when human assessors judge relevance using the usual 4-levels scale for Highly relevant, Relevant, Marginally relevant, Not relevant, it is common to translate them into the numeric values 3, 2, 1, and 0. Turning to system outputs, of course the Retrieval Status Values are an assignment; but any ranked list of retrieved documents can easily be converted to an assignment, for example using the reciprocal of the rank. If the abstract task is not Ranking but, let us say, Clustering, the gold and system output will be again an assignment of numbers to documents, with the natural convention that two documents having the same value means that they are in the same cluster, according to the gold and/or the system output; similarly for Classification.

On these basis we define a metric as a function that, given a system output σ and a gold α , returns a real value that depends on how much σ is *close* to α .

Definition 6 (*Metric*) A metric is a function $\mathcal{M} : \Omega^2 \longrightarrow \mathbb{R}$.

We remark that some authors and metrics require a bounded codomain for \mathcal{M} . Mofat’s (2013) first property is Boundedness (see Sect. 2.3). Several metrics assume values in $[0, 1]$. However, this is not always the case: some metrics (that we will analyze in the following) such as Utility metrics (for classification) are unbounded and assume values in $(-\infty, +\infty)$, other metrics like DCG (for IR), and MAE (for quantitation) in $[0, +\infty)$, and Pearson and Spearman in $[-1, +1]$. We choose the most general possibility for two reasons: (i) we aim at a general framework, thus we do not want to exclude metrics and to do so we focus on monotonicity and invariance properties; and (ii) normalizations can be applied, though this seems a technical and minor issue that we leave for future work.

However, we have defined two notions of closeness; therefore we define two families of metrics, as follows.

4.2 First family: value-oriented metrics

Value-oriented metrics quantify to what extent a system resembles the values assigned to items in the gold. The more the system values are close to the gold values, the more the metric returns high scores for the system. For instance, the values spam/non spam assigned by a spam filter should match with the values given by the human references. Given that the concept of closeness defined in the previous section depends on the scale type, the definition of a value-oriented metric is also dependent on the scale type. In addition, transforming both the system output and the gold by the same permissible transformation function should not affect the metric result. For instance, we can apply a bijective transformation from the value representing the category “spam” to the value representing the category “trash-messages” (that, being bijective, is in \mathcal{F}_N , i.e., is a permissible transformation function for the nominal scale) as long as we do it for both the system output and the gold.

We first define the following two properties.

Property 1 (VOI, value-oriented invariance) *A metric \mathcal{M} is value-oriented invariant for the scale type T if for any reference gold α and system output σ , both of them on the same set of objects \mathcal{D} , the metric value does not change by applying the same permissible transformation to both α and σ :*

$$\forall \alpha, \sigma \in \Omega, \forall f \in \mathcal{F}_T \left(\mathcal{M}(\sigma, \alpha) = \mathcal{M}(f(\sigma), f(\alpha)) \right). \tag{13}$$

Property 2 (VOM, value-oriented monotonicity) *A metric \mathcal{M} is value-oriented monotonic for the scale type T if for any reference gold α and system outputs σ and σ' , all three of them on the same set of objects \mathcal{D} , it holds that if σ is value-closer to α than σ' , then the metric value for σ has to be higher than that for σ' . In formulas:*

$$\forall \alpha, \sigma, \sigma' \in \Omega \left(\sigma \triangleleft_T^\alpha \sigma' \implies \mathcal{M}(\sigma, \alpha) > \mathcal{M}(\sigma', \alpha) \right). \tag{14}$$

We can now specialize Definition 6 and formally define a value-oriented metric on the basis of value-oriented assignment closeness (Definition 3) as follows.

Definition 7 (*Value-oriented metric*) *A value-oriented evaluation metric for the scale type T is a metric that satisfies the two properties VOI and VOM for the scale type T .*

Therefore, the metrics of this family need to satisfy just two basic properties: (i) invariance to permissible transformation functions (VOI), i.e., the metric value does not change when transforming both assignments in the same permissible way, and

(ii) monotonicity to value-oriented closeness (VOM), i.e., if an assignment is value-closer to the gold than another, then the former has a higher metric value. Note that the two

properties VOI and VOM depend on the scale type: a given function could be a metric for a scale type T and not be a metric for another scale type T' .⁵

4.3 Second family: equivalence-oriented metrics

Directly comparing the values of system outputs to the values assigned by the gold is in some cases too strict. For instance, the purpose of search engines consists of offering the most relevant documents rather than quantifying the relevance of documents: the key point is that any assignment that keeps the ranking of documents in the search engine output is equally effective. That is, any assignment in the same equivalence class for the ordinal scale type.

The metrics of the second family, the equivalence-oriented metrics, are formally defined on the basis of the following two properties.

Property 3 (EOI, equivalence-oriented invariance) *A metric \mathcal{M} is equivalence-oriented invariant for the scale type T if for any reference gold α and system output σ , both of them on the same set of objects \mathcal{D} , the metric value does not change by applying any permissible transformation to σ . In formulas:*

$$\forall \alpha, \sigma \in \Omega, \forall f \in \mathcal{F}_T \left(\mathcal{M}(\sigma, \alpha) = \mathcal{M}(f(\sigma), \alpha) \right). \quad (15)$$

Property 4 (EOM, equivalence-oriented monotonicity) *A metric \mathcal{M} is equivalence-oriented monotonic for the scale type T if for any reference gold α and system outputs σ and σ' , all three of them on the same set of objects \mathcal{D} , it holds that if σ is equivalence-closer to α than σ' , then the metric value for σ has to be higher than that for σ' . In formulas:*

$$\forall \alpha, \sigma, \sigma' \in \Omega \left(\sigma \sqsubset_T^\alpha \sigma' \implies \mathcal{M}(\sigma, \alpha) > \mathcal{M}(\sigma', \alpha) \right). \quad (16)$$

We can now specialize again Definition 6 and define an equivalence-oriented metric on the basis of equivalence-oriented assignment closeness (Definition 4) as follows.

Definition 8 (*Equivalence-oriented metric*) *An equivalence-oriented evaluation metric for the scale type T is a metric that satisfies the two properties EOI and EOM for the scale type T .*

Therefore, also the metrics of this family need to satisfy two basic properties only, namely invariance (EOI) and monotonicity (EOM), although these are defined in a slightly different way from the corresponding properties of the previous family of metrics (VOI and VOM): in EOI the function f is applied to σ only, and in EOM equivalence-oriented assignment closeness is used. As above, we will use the scale type as a subscript for the EOI and EOM properties when needed (see Footnote 5).

⁵ In the following, we use the specific scale type as a subscript of the property name to specify the property for that scale type when needed. So, for example, VOI_n is the VOI property for the nominal scale type.

Table 7 The eight classes of metrics

		Family	
		Value-Oriented (VOI + VOM)	Equivalence-Oriented (EOI + EOM)
Scale type	N	1. Classification	2. Clustering
	O	3. Ordinal Classification	4. Ranking
	I	5. Quantitation-1	6. Quantitation-3
	R	7. Quantitation-2	8. Quantitation-4

4.4 The example revisited

We provide some intuition by discussing again the example in Sect. 3.7 and Table 6: here ρ is the gold (α using the notation of this section) and the ω_i are the system outputs (σ_i). Let us first interpret the assignments as being of the I scale, and focus on ω_1 and ω_2 . Since ω_2 is value-closer to ρ than ω_1 ($\omega_2 \triangleleft_I^{\rho} \omega_1$, see Formula (10)), a value-oriented metric will assign a higher value to ω_2 than to ω_1 ($\mathcal{M}(\omega_2, \rho) > \mathcal{M}(\omega_1, \rho)$), because of the VOM property. Conversely, since ω_1 is equivalence-closer to ρ than ω_2 ($\omega_1 \sqsubset_I^{\rho} \omega_2$, see Formula (11)), an equivalence-oriented metric will assign a higher value to ω_1 than to ω_2 ($\mathcal{M}(\omega_1, \rho) > \mathcal{M}(\omega_2, \rho)$), because of the EOM property. Metrics of the first family reward ω_2 for “almost guessing” the correct ρ values; metrics of the second family reward ω_1 for guessing better the ratios between the intervals (i.e., the meaningful statements for I) among the values in ρ .

Indeed, ω_1 guess of the ratios of the intervals is not only better, it is perfect. This can be seen also considering the EOI property: if the values in ω_1 are multiplied by ten (a permissible transformation function for I) we obtain exactly ρ . Since it is not possible to do better than ω_1 , an equivalence-oriented metric should assign to ω_1 a higher value than any other one. Note that the above remarks hold also when replacing ω_1 with ω_4 (apart from changing the permissible transformation function), as ω_4 is equivalent to ω_1 (simply use the transformation $f(x) = \frac{x-2}{10}$) and to ρ ($f(x) = x - 2$). This also means, again because of EOI, that $\mathcal{M}(\omega_1, \rho) = \mathcal{M}(\omega_4, \rho)$.

Changing the scale will in general change the situation. For example, if we now interpret the same assignments as being of the R scale, we get different outcomes. On the R scale, ω_4 is not equivalent to ω_1 and ρ anymore: there is no function in \mathcal{F}_R that maps the one into the other two. Indeed, since ω_1 is equivalence-closer to ρ than ω_4 (see Formula (12)), equivalence-oriented metrics for the ratio scale will assign a higher value to ω_1 than to ω_4 ($\mathcal{M}(\omega_1, \rho) > \mathcal{M}(\omega_4, \rho)$), because of the EOM property.

4.5 Eight classes of metrics

Our framework is now complete: we have provided two definitions, one for each metric family: value- and equivalence-oriented metric. Both definitions can be applied at different scale types: nominal, ordinal, interval or ratio. By combining the four scale types N, O, I, R and the two families of metrics (value- and equivalence-oriented) we obtain the *eight classes of metrics* summarized in Table 7.

In the following we prove some theorems that show that:

- The basic axioms proposed in the literature for specific abstract tasks can be derived from the general metric definition, taking into account the particular combination of family and scale type. Notice that we have used the term *basic axiom* instead of *axiom*. The reason is that, as we have seen in Sect. 2, some axioms depend on the particular task, while other (basic) axioms are common for any task that fits into the corresponding abstract task.
- Existing abstract tasks, and the corresponding metrics, actually fit into our classification. More specifically, we show that each information access abstract task (classification, clustering, etc.) corresponds to a metric class, i.e., a particular combination of metric family and scale type, as well as that metrics that fit in the same category according to these two dimensions are used in the literature for that abstract task.
- The theoretical limitations of metrics that have been identified in the literature (i.e., metrics that do not satisfy basic axioms) can be explained also in the general framework proposed in this paper.
- By exploring the classes of metrics along the two dimensions (scale type and family of metric), it is possible to address evaluation gaps and provide formal definitions, for example, of Ordinal Classification metrics, which have not been addressed yet.

5 Properties and scale types

We start by making explicit some implication relationships between the four properties VOI, VOM, EOI, and EOM at different scale types. These are derived from the fact that permissible transformation functions are subsumed across scale types (see, in Appendix A, Sect. A.2 and in particular Formulas (30) and (31)). That is, the set of bijective functions includes the set of monotonic functions, which in turn includes the set of linear affinity functions. Therefore, closeness for low scale types (e.g., nominal) implies closeness for higher scale types (e.g., interval). The resulting relationships are listed in the following lemma and in the two subsequent corollaries, with the aims of: (i) provide the basis to prove some properties in the following of the paper, and (ii) help to better understand the meaning of the four axioms VOI, VOM, EOI, EOM and their relationships with the four scales \mathbb{N} , \mathbb{O} , \mathbb{I} , \mathbb{R} .

Lemma 2 (Four properties, four scale types) *The following relationships hold among the four properties (VOI, VOM, EOI, EOM) and the four scale types (\mathbb{N} , \mathbb{O} , \mathbb{I} , \mathbb{R}).*

- (a) *If a metric satisfies VOI for a certain scale type, then it satisfies VOI for higher scale types:*

$$\forall T, T' \in \{\mathbb{N}, \mathbb{O}, \mathbb{I}, \mathbb{R}\}, T < T' \quad (\text{VOI}_T \implies \text{VOI}_{T'}).$$

- (b) *If a metric satisfies EOI for a certain scale type, then it satisfies EOI for higher scale types:*

$$\forall T, T' \in \{N, O, I, R\}, T < T' (EOI_T \implies EOI_{T'}).$$

- (c) *VOI for the nominal or ordinal scale type and VOM for higher scale types are incompatible:*

$$\forall T \in \{N, O\}, T' \in \{O, I, R\}, T < T' (\neg(VOI_T \wedge VOM_{T'})).$$

- (d) *VOI for the nominal or ordinal scale type and EOM for higher scale types are incompatible:*

$$\forall T \in \{N, O\}, T' \in \{O, I, R\}, T < T' (\neg(VOI_T \wedge EOM_{T'})).$$

- (e) *VOM and EOM are incompatible, whatever the scale type:*

$$\forall T, T' \in \{N, O, I, R\} (\neg(VOM_T \wedge EOM_{T'})).$$

- (f) *VOM_I and VOM_R are equivalent:*⁶

$$VOM_I \iff VOM_R.$$

- (g) *VOM and EOI are incompatible, whatever the scale type:*

$$\forall T, T' \in \{N, O, I, R\} (\neg(VOM_T \wedge EOI_{T'})).$$

- (h) *EOM for a certain scale type and EOI for lower scale types are incompatible:*

$$\forall T, T' \in \{N, O, I, R\}, T' < T (\neg(EOM_T \wedge EOI_{T'})).$$

From this lemma, we can infer the following corollaries. It is easy to see that items (c), (d), (e), (g), and (h) can be restated as implications.

Corollary 1 (Incompatibilities and implications) *Items (c) and (d) of Lemma 2 can be restated as follows. If a metric satisfies VOI for the nominal or ordinal scale type, then it does not satisfy VOM and EOM for higher scale types, and vice-versa if a metric satisfies VOM or EOM for higher scale types, then it does not satisfy VOI for the nominal or ordinal scale type:*

$$\forall T \in \{N, O\}, T < T' (VOI_T \implies \neg(VOM_{T'})) \tag{17}$$

$$\forall T \in \{N, O\}, T < T' (VOM_{T'} \implies \neg(VOI_T)) \tag{18}$$

$$\forall T \in \{N, O\}, T < T' (VOI_T \implies \neg(EOM_{T'})) \tag{19}$$

$$\forall T \in \{N, O\}, T < T' (EOM_{T'} \implies \neg(VOI_T)). \tag{20}$$

⁶ This property is not used anywhere in this paper; we include it for completeness.

Item (e) of the lemma can be restated as follows. If a metric satisfies VOM for a certain scale type, then it does not satisfy EOM for any scale type, and vice-versa:

$$\forall T, T' \in \{N, O, I, R\} (VOM_T \implies \neg(EOM_{T'})) \quad (21)$$

$$\forall T, T' \in \{N, O, I, R\} (EOM_{T'} \implies \neg(VOM_T)). \quad (22)$$

Item (g) of the lemma can be restated as follows. If a metric satisfies EOI for any scale type, then it does not satisfy VOM for any scale type, and vice-versa:

$$\forall T, T' \in \{N, O, I, R\} (EOI_T \implies \neg(VOM_{T'})) \quad (23)$$

$$\forall T, T' \in \{N, O, I, R\} (VOM_{T'} \implies \neg(EOI_T)). \quad (24)$$

Item (h) of the lemma can be restated as follows. If a metric satisfies EOM for a certain scale type, then it does not satisfy EOI for lower scale types, and vice-versa:

$$\forall T, T' \in \{N, O, I, R\}, T' < T (EOM_T \implies \neg(EOI_{T'})) \quad (25)$$

$$\forall T, T' \in \{N, O, I, R\}, T' < T (EOI_{T'} \implies \neg(EOM_T)). \quad (26)$$

Another result is that, knowing that a metric fits into one metric class ensures that that metric does not fit into other definitions, with only one exception. The unification between value-oriented metrics for interval and ratio scale types is due to the fact that the concepts of closeness for those two scale types are equivalent (see Table 5).

Corollary 2 (Compatibility of value-oriented interval and ratio) *Under our definition of metrics, there exists only one case in which a metric can be classified in more than one class (see Table 7): the value-oriented metrics for the interval and ratio scale types (5 and 7 in the table).*

6 Basic axioms in the literature

In this section we state some theorems that prove that the basic axioms proposed in the literature for classification, clustering, and ranking, i.e., GMON (Generalized Strict Monotonicity Axiom), GHC (Generalized Homogeneity / Completeness), and PRI (Priority Axiom) (see Sect. 2) can be derived in our framework. As anticipated in Sect. 4.5, we focus on the *basic* axioms that we have identified in Sect. 2 (i.e., the axioms marked with (*) and shown in italics in the tables in Sect. 2), leaving aside the task-dependent axioms. As noted in Sect. 2.4, there are no axiomatics, and no basic axioms, for quantitation, at least in the context of information access evaluation: it is left out from this section, and analyzed in Sect. 8.1.

6.1 Classification: GMON is equivalent to VOM_N

As discussed in Sect. 2.1, the common axiom that can be applied to any classification task is the Strict Monotonicity Axiom (MON). It states that if σ and σ' are two classifiers and α is the ground truth, all of them over a set of documents \mathcal{D} , and if σ and σ' differ only for a single document $d \in \mathcal{D}$, which is correctly classified for σ and wrongly for σ' , then the metric value must be higher for σ . More formally, if

$$\exists! d \in \mathcal{D} \left(\forall d' \in \mathcal{D} \setminus \{d\} (\sigma(d') = \sigma'(d')) \wedge \alpha(d) = \sigma(d) \neq \sigma'(d) \right)$$

then $\mathcal{M}(\sigma, \alpha) > \mathcal{M}(\sigma', \alpha)$.

This definition requires that σ and σ' return the same result for all documents different from d . However, this is not strictly necessary: we can slightly generalise MON requiring that both systems are accurate or wrong, with respect to the gold, for the same documents, except for d .

Axiom 1 (GMON, generalized strict monotonicity axiom) *Let α, σ , and σ' be three assignments. If every document with an error in σ is also an error in σ' , and there exist an error in σ' which is not an error in σ , i.e.,*

$$\begin{aligned} \forall d \in \mathcal{D} (\alpha(d) = \sigma(d) \vee \alpha(d) \neq \sigma'(d)) \\ \exists d \in \mathcal{D} (\alpha(d) = \sigma(d) \neq \sigma'(d)) \end{aligned}$$

then the metric value must be higher for σ than σ' :

$$\mathcal{M}(\sigma, \alpha) > \mathcal{M}(\sigma', \alpha).$$

When there are only two classes (only two possible values in σ and α), GMON and MON are equivalent, given that if $\sigma(d) \neq \alpha(d)$ and $\sigma'(d) \neq \alpha(d)$ then $\sigma(d) = \sigma'(d)$.

We can now prove the following theorem.

Theorem 1 (VOM_N and GMON) *The VOM property for the nominal scale type (VOM_N) and the Generalized Strict Monotonicity (GMON) axiom are equivalent.*

6.2 Clustering: GHC is equivalent to EOM_N

As mentioned in Sect. 2.2, the basic clustering axioms Homogeneity and Completeness can be generalized into a unique GHC axiom, which can be formalized as follows.

Axiom 2 (GHC, generalized homogeneity/completeness) *Let α, σ , and σ' be three assignments. If (i) for each document pair if σ' is correct then also σ is correct, i.e.,*

$$\begin{aligned} \forall d_i, d_j \in \mathcal{D} \left((\sigma'(d_i) = \sigma'(d_j) \wedge \alpha(d_i) = \alpha(d_j) \implies \sigma(d_i) = \sigma(d_j)) \wedge \right. \\ \left. (\sigma'(d_i) \neq \sigma'(d_j) \wedge \alpha(d_i) \neq \alpha(d_j) \implies \sigma(d_i) \neq \sigma(d_j)) \right), \end{aligned} \tag{27}$$

and (ii) there exists at least a document pair d_1, d_2 such that σ adds to σ' a correct relation, i.e., σ is correct and σ' is not, i.e.,

$$\exists d_1, d_2 \in \mathcal{D} \left((\sigma'(d_1) = \sigma'(d_2) \wedge \alpha(d_1) \neq \alpha(d_2) \wedge \sigma(d_1) \neq \sigma(d_2)) \vee (\sigma'(d_1) \neq \sigma'(d_2) \wedge \alpha(d_1) = \alpha(d_2) \wedge \sigma(d_1) = \sigma(d_2)) \right) \quad (28)$$

then the metric value must be higher for σ than σ' :

$$\mathcal{M}(\sigma, \alpha) > \mathcal{M}(\sigma', \alpha).$$

The following theorem states that the GHC and the EOM properties for the nominal scale type are equivalent.

Theorem 2 (EOM_N and GHC) *The EOM property for the nominal scale type (EOM_N) and the Generalized Homogeneity and Completeness (GHC) axiom are equivalent.*

6.3 Ranking: PRI is equivalent to EOM₀

We have seen in Sect. 2.3 that the basic axiom *Swapping* appears in most axiomatics and that it can be generalized as the Priority axiom (PRI). PRI states that swapping two contiguous documents in the ranking according to the gold necessarily increases the score. We formalize it as follows.

Axiom 3 (PRI, priority axiom) *Let α , σ , and σ' be three assignments such that d_i and d_j have contiguous values at scale type 0 in both σ and σ' . If:*⁷

$$\exists i, j \left(\alpha(d_i) > \alpha(d_j) \wedge \sigma(d_i) > \sigma(d_j) \wedge \sigma'(d_i) < \sigma'(d_j) \wedge \forall k, l \neq i, j (\sigma(d_k) > \sigma(d_l) \Leftrightarrow \sigma'(d_k) > \sigma'(d_l)) \right) \quad (29)$$

then the metric value must be higher for σ than σ' :

$$\mathcal{M}(\sigma, \alpha) > \mathcal{M}(\sigma', \alpha).$$

We can prove the following theorem.

Theorem 3 (EOM₀ and PRI) *The EOM property for ordinal scale type (EOM₀) and the Priority (PRI) axiom are equivalent.*

In other words, our definition of metrics captures the basic axiom for metrics in ranking tasks. We now turn to analyse the implications for specific tasks and metrics.

⁷ Even when the scale type of reference is 0, by $\sigma(d) > \sigma(d')$ we mean that for the σ assignment d is more relevant than d' .

Table 8 Metric analysis for classification, clustering, and ranking abstract tasks

	Classification metrics		Clustering metrics		Ranking metrics	
	Accuracy, MAAC, Phi	<i>F-measure, Odds, Lam%</i>	Entropy, BCubed, Count pairs, Mutual Inf.	<i>Purity and Inv. Purity, F-measure</i>	MAP, ERR, nDCG, RBP, Kendall, Spearman	<i>P@N, R@N, MRR</i>
VOI _N	●	○	–	–	× ^c	–
VOM _N	●	×	× ^g	–	× ^g	× ^g
VOI ₀	○ ^a	○ ^a	–	–	–	–
VOM ₀	× ^c	× ^c	× ^g	× ^g	× ^g	× ^g
VOI _I	○ ^a	○ ^a	–	–	–	–
VOM _I	× ^c	× ^c	× ^g	× ^g	× ^g	× ^g
VOI _R	○ ^a	○ ^a	–	–	–	–
VOM _R	× ^c	× ^c	× ^g	× ^g	× ^g	× ^g
EOI _N	× ^g	–	●	○	× ^h	–
EOM _N	× ^e	–	●	×	–	–
EOI ₀	× ^g	–	○ ^b	○ ^b	●	○
EOM ₀	× ^e	× ^d	× ^h	× ^h	●	×
EOI _I	× ^g	–	○ ^b	○ ^b	○ ^b	○ ^b
EOM _I	× ^e	× ^d	× ^h	× ^h	× ^h	× ^h
EOI _R	× ^g	–	○ ^b	○ ^b	○ ^b	○ ^b
EOM _R	× ^e	× ^d	× ^h	× ^h	× ^h	× ^h

Crosses (×) indicate that the property is not satisfied, circles (○) that it is satisfied and filled circles (●) emphasize that both properties are satisfied. The superscript letter represents the item of Lemma 2 that supports the conclusion. The dash (–) indicates that the property is not formally checked in this paper (we leave them for future work)

7 Metrics analysis

In this section, we analyse existing metrics in terms of our theoretical framework. We have the twofold aim of: (i) showing that existing abstract tasks and corresponding metrics are explained by our framework, and (ii) that the theoretical limitations of metrics that have been identified in the literature (i.e., metrics that do not satisfy basic axioms) are captured in our framework as well. As noted in the previous section, we postpone the quantitation case to Sect. 8.1.

Table 8 summarises the analysis carried out in this section. The columns represent the metrics categorised by abstract tasks. The rows represent the properties VOI, VOM, EOI, and EOM, at different scale types. Circles indicate that properties are satisfied, black circles emphasize that both properties (from the corresponding definition of evaluation metric) are satisfied. As the table shows, only metrics in the corresponding category satisfy our definition of metric. In addition, for each abstract task category, there exist metrics which are not able to satisfy both properties. This does not mean that these metrics are absolutely useless. For instance, Purity and Inverse Purity in clustering, or P@N in IR have the advantage of being easy to interpret. However, the evaluation results have to be analysed carefully to prevent misinterpretations of systems’ quality. We will see in this section that these cases correspond with theoretical metric drawbacks previously reported in the literature with practical implications.

7.1 Classification: value-oriented, nominal

We define classification metrics as follows.

Definition 9 (*Classification metric*) A classification metric is a value-oriented metric for the nominal scale type.

According to our aim (i) above, we want to show that metrics satisfying VOI and VOM for the nominal scale type are those that are used for classification problems in the literature. Most of the metrics used in classification tasks are combinations of the contingency matrix components, i.e., the amount of true and false, positive and negative samples. Formally, the contingency matrix can be defined as follows.

Definition 10 (*Contingency matrix*) Being the system output σ and the gold α two functions over a limited amount of values \mathcal{V} , the contingency matrix $C : \mathcal{V} \times \mathcal{V} \rightarrow \mathcal{N}$ is:

$$C_{\sigma,\alpha}(x,y) = \text{card}(\{d \in \mathcal{D} \mid \sigma(d) = x \wedge \alpha(d) = y\}).$$

We can prove easily that $C(x, y)$ is invariant across the permissible transformation functions for the nominal scale type $\mathcal{F}_{\mathbb{N}}$ (i.e., the bijective functions, see Sect. A.2 in Appendix A) applied to σ and α . In other words, changing the names of categories without merging them (as it is done with bijective functions) in both the gold and the system output does not affect the contingency table. That is, being f_b any bijective function

$$C_{\sigma,\alpha}(x,y) = C_{f_b(\sigma),f_b(\alpha)}(f_b(x),f_b(y)).$$

Therefore, we can state the following theorem.

Theorem 4 (VOI _{\mathbb{N}} and contingency matrix) *Any function over the elements in the contingency matrix satisfies the VOI property for the nominal scale type (VOI _{\mathbb{N}}).*

Table 8 includes some metrics used in classification tasks, such as Accuracy, Macro-Average Accuracy, F-measure, Odds ratio, Lam%. All of them are computed from the contingency matrix. Therefore, according to the previous theorem, they satisfy VOI _{\mathbb{N}} . According to Lemma 2(a) they also satisfy VOI for the rest of (higher) scale types.

We can prove that some common metrics applied in classification satisfy VOM _{\mathbb{N}} (and, given Theorem 1, GMON).

Theorem 5 (VOM _{\mathbb{N}} and classification metrics) *The metrics Accuracy, Macro Average Accuracy and Phi Correlation satisfy the VOM property for the nominal scale type (VOM _{\mathbb{N}}).*

Given that these metrics also satisfy VOI _{\mathbb{N}} (according to Theorem 4), we can state that they fit into our definition of value-oriented metrics for the nominal scale type, and therefore, they are classification metrics.

We now turn to our aim (ii) above. The main theoretical drawbacks of classification metrics are related with MON. For instance, according to Sebastiani (2015), the F-measure (computed as the harmonic mean of precision and recall) does not satisfy MON when there is a zero value in one component of the contingency matrix. According to Qi et al. (2010), the classification metric Lam% has the same drawback when

zero values appear in the contingency matrix. Something similar happens with the Odds ratio. This problem can be solved by considering the contingency table as a probabilistic distribution and applying smoothing techniques (Amigó et al. 2018), but this is not the focus of this paper. Given that they do not satisfy MON, they do not satisfy also GMON, and therefore, VOM_N . Mutual Information (MI) is another metric used in classification. However, we can assert that it does not satisfy GMON (and VOM_N), given that according to MI, an output achieves the highest score even if the label names are replaced. We will see shortly that MI is a clustering metric.

The second and third columns of Table 8 summarize the properties satisfied by these metrics according to the implications derived from Lemma 2. In general, the main result of this analysis is that metrics used in classification tasks fit into our definition, while the main limitations of metrics such as F-measure, Lam% or Odds identified in the literature are also captured by our model.

7.2 Clustering: equivalence-oriented, nominal

We define clustering metrics as follows.

Definition 11 (*Clustering metric*) A clustering metric is an equivalence-oriented metric for the nominal scale type.

We aim to show that this definition actually captures the existing clustering metrics and their desirable basic constraints. We start by noting that existing clustering metrics such as Purity and Inverse Purity, BCubed precision and Recall, Entropy and Class Entropy, F-measure, or clustering metrics based on counting, are all functions over a set partition. A partition is the standard set concept: a decomposition of a set in subsets such that their (pairwise) intersection is empty and their union is the original set. A function over a partition is any function that takes the original set and its partition as input parameters. We can now state as a theorem that all the above metrics satisfy EOI_N .

Theorem 6 (EOI_N and functions over a partition) *Any function over a set partition satisfies the EOI property for the nominal scale type (EOI_N).*

Checking if each metric proposed in the literature satisfies EOM_N (or GHC, which is equivalent according to Theorem 2), is too complex to be included in this paper. For this reason, we will analyse the existing metrics by using the categorisation proposed by Amigó et al. (2009). Then, we will check to what extent a category can produce metrics that satisfy EOM and EOI at the nominal scale type.

A first category is called *counting pairs* based metrics [e.g., Rand Statistic, Jaccard Coefficient, or Folkes and Mallows (Meila 2003; Halkidi et al. 2001)], that count how many pairs correspond to the gold in terms of same/different cluster: correct relationships between pairs of items increases the score. This principle matches directly with the GHC conditions, which is equivalent to EOM_N . On the other hand, the metric BCubed (Amigó et al. 2009) also increases with the amount of document pairs that are consistent with the gold. It also satisfies GHC. Therefore, according to Theorem 2, we can state that metrics based on counting pairs fit into our definition of clustering metric.

A second category is the *entropy based* metrics. Some examples are Entropy and Class Entropy (Wu et al. 2003), Variation of Information (Meila 2003), Mutual Information (Xu et al. 2003) or V-measure (Rosenberg and Hirschberg 2007). Proving that all those metrics satisfy EOM and EOI requires too much analysis for this paper: we focus on Entropy and Class entropy only. Given a set of documents \mathcal{D} , a ground truth clustering α and a clustering σ , the average Entropy of clusters is computed as (Wu et al. 2003)

$$E(\sigma, \alpha) = - \sum_{c \in \mathcal{V}(\sigma)} \left(P(\sigma(d) = c) \cdot \sum_{l \in \mathcal{V}(\alpha)} \left(P(\alpha(d) = l \mid \sigma(d) = c) \cdot \log(P(\alpha(d) = l \mid \sigma(d) = c)) \right) \right),$$

where the sums are over the documents $d \in \mathcal{D}$, as well as the probability P computed by frequency count as usual, $\mathcal{V}(\alpha)$ is the set of different values generated by the gold assignment α , and $\mathcal{V}(\sigma)$ is the set of values generated by the system assignment σ . The Class Entropy is defined as

$$CE(\sigma, \alpha) = - \sum_{l \in \mathcal{V}(\alpha)} \left(P(\alpha(d) = l) \cdot \sum_{c \in \mathcal{V}(\sigma)} \left(P(\sigma(d) = c \mid \alpha(d) = l) \cdot \log(P(\sigma(d) = c \mid \alpha(d) = l)) \right) \right).$$

Notice that the evaluation score is inversely correlated with the entropy values. Then we can prove the following theorem.

Theorem 7 (EOM_N and entropy metrics) *Entropy and Class Entropy satisfy the EOM property for the nominal scale type (EOM_N).*

The conclusion is that BCubed, metrics based on counting pairs, and entropy based metrics are able to satisfy EOI_N and EOM_N. Therefore, they fit into our definition of clustering metric. However, not all metrics used in clustering tasks fit into our definition. In particular, metrics based on set matching such as F-measure [notice that F-measure has a different meaning in the context of clustering (Amigó et al. 2009)] or Purity and Inverse Purity do not satisfy Completeness (Amigó et al. 2009). Therefore, they do not satisfy GHC and, according to Theorem 2, neither EOM_N. The reason is that they assume a certain correspondence between system output and gold clusters. This produces a lack of sensitivity in some cases.

The fourth and fifth columns in Table 8 illustrate the properties satisfied by these metrics according to the implications derived from Lemma 2. In summary, we can say that most metrics used in clustering fit into our definition, capturing the limitations described in the literature.

As we mentioned before, there exist other axioms that are not captured by our definition, but they are task-dependent. For instance, the range of values (Meila 2003), the ability to join single clusters into a rag bag cluster (Amigó et al. 2009), or the robustness under the overweighting of big clusters due to the combinatory explosion of document pairs.

7.3 Ranking: equivalence-oriented, ordinal

We define ranking metrics as follows.

Definition 12 (*Ranking metric*) A ranking metric is an equivalence-oriented metric for the ordinal scale type.

Let us see that ranking metrics satisfy EOI_0 and EOM_0 . Let us first consider the obvious fact that ranking metrics compare system output ranking against the gold. By definition, a ranking is invariant across monotonic functions (the permissible transformation functions for ordinal scale types, \mathcal{F}_0). Therefore, we can state the following theorem.

Theorem 8 (EOI_0 and ranking metrics) *Metrics that compare rankings with a gold standard satisfy the EOI property for the ordinal scale type (EOI_0).*

Concerning EOM_0 , we have seen in Theorem 3 that it is equivalent to the Priority axiom (PRI). According to Amigó et al. (2013), many metrics applied in ranking problems actually satisfy PRI. Therefore, we can state the following theorem.

Theorem 9 (EOM_0 and ranking metrics) *The metrics MAP, DCG, nDCG, RBP, ERR, and ordinal correlation coefficients such as Kendall or Spearman satisfy the EOM property for the ordinal scale type (EOM_0).*

Therefore, many of the metrics used in ranking problems actually fit our definition. This is the case of metrics such as MAP, DCG, nDCG, RBP, ERR and also ordinal correlation coefficients such as Kendall or Spearman. However, these last two coefficients are normally not used in IR, since the collection contains a huge amount of documents that will never be explored by the user. Indeed, IR evaluation metrics, besides satisfying the priority axiom, also give more weight to the top of the ranking returned by the system. This is captured for example by Moffat's properties Convergence and Top-weightedness (2013), or by Amigó et al.'s Deepness, Closeness Threshold, and Deepness Threshold (2013) (see Table 3). These properties are not captured in our framework and have to be added explicitly if needed; this is usually the case, although, as we already discussed in Sect. 2.3, one might imagine a ranking task where the top-weightedness property is undesirable: for instance, if we need to rank a set of documents, and we know that the user will explore all the documents anyway.

However, not every metric used in ranking satisfies the Priority axiom (Amigó et al. 2013). This is the case of some metrics such as Precision at N, Recall at N or Maximum Reciprocal Rank: $P@N$ and $R@N$ do not consider the order of documents before position N, and MRR does not consider the order of documents after the first relevant one. Therefore, according to Theorem 3 we can infer that they do not satisfy EOM_0 : they do not fit into our definition of ranking metrics. Table 8 (last two columns) illustrates the ranking metrics.

In summary, again, most of the metrics used in ranking problems fit into our definition, with some exceptions whose limitations have been discussed in the literature.

8 Other tasks and metrics

The previous two sections show how the classical abstract tasks (classification, clustering, and ranking) and their metrics can be modeled in our framework. In this section we aim to demonstrate the generality of the framework, showing how it adapts also to other information access tasks, by (i) analyzing the metrics for the quantitation, and (ii) showing that the framework leads to ordinal classification, which has not yet been studied from an axiomatic perspective.

8.1 Quantitation: value- and equivalence-oriented, interval and ratio

The proposed framework gives us the opportunity of modeling tasks at the higher scale types I and R. We now analyze the four quantitation variants shown in Table 7.

8.1.1 Quantitation-1: Value-oriented, interval

Let us analyse the effect of applying the metric definition for the interval scale type. For instance, a value-oriented metric for the interval scale type must be invariant under linear transformations, and it must increase when every assignment is value-closer to the gold-standard. Let us consider the widely used Mean Absolute Error (MAE). It is computed as the average difference between the σ and α values:

$$\text{MAE}(\sigma, \alpha) = \text{Avg}_{d \in \mathcal{D}} |\alpha(d) - \sigma(d)|.$$

Notice that the definition of this metric directly matches with the definition of closeness for these scale types. This metric satisfies VOM for both interval and ratio scale types. However, as it is, it does not fit into the definition of value-oriented metric, since VOI_I (as well VOI_R) does not hold. For instance, transforming both the assignments by multiplying them by a constant factor (a permissible transformation function) affects the average difference. However, the average error has always to be expressed in terms of a unit. For instance, a MAE of 2 when measuring temperature has no sense: one should say a MAE of 2 centigrades. That is, we need to incorporate a unit $|\alpha(d_0) - \alpha(d'_0)|$ which depends of empirical observations over a fixed pair of objects (e.g., a centigrade is a hundredth of the temperature difference between ice and water vapor). Then, applying a transformation also implies transforming the unit, and in this way the average error is invariant for the interval scale type. We can define MAE with a reference difference as:

$$\text{MAE}_{\text{RD}}(\sigma, \alpha) = \text{Avg}_{d \in \mathcal{D}} \left(\frac{|\alpha(d) - \sigma(d)|}{|\alpha(d_0) - \alpha(d'_0)|} \right).$$

After this definition, we can state the following theorem.

Theorem 10 (VOI_I , VOM_I and MAE) *The Mean Absolute Error with a reference difference is a value-oriented metric for the interval scale type, i.e., MAE_{RD} satisfies VOI_I and VOM_I .*

Table 9 Metric analysis for high scale types (I and R) and value-oriented for ordinal scale type

	Quantitation metrics				
	MAE	MAE _{RD}	MAE _{RU}	CORR	COS
VOI _N	x ^c	x ^c	x ^c	x ^c	x ^c
VOM _N	–	–	–	x ^g	x ^g
VOI ₀	x ^c	x ^c	x ^c	x ^c	x ^c
VOM ₀	–	–	–	x ^g	x ^g
VOI _I	x	●	x	–	–
VOM _I	○	●	○	x ^g	x ^g
VOI _R	x	●	●	–	–
VOM _R	○	●	●	x ^g	x ^g
EOI _N	x ^g	x ^g	x ^g	x ^h	x ^h
EOM _N	x ^e	x ^e	x ^e	–	–
EOI ₀	x ^g	x ^g	x ^g	x ^h	x ^h
EOM ₀	x ^e	x ^e	x ^e	–	–
EOI _I	x ^g	x ^g	x ^g	●	x ^h
EOM _I	x ^e	x ^e	x ^e	●	–
EOI _R	x ^g	x ^g	x ^g	○ ^b	●
EOM _R	x ^e	x ^e	x ^e	x ^h	●

Crosses (x), circles (○), filled circles (●), dashes (–) and the super-script letters have the same meaning as in Table 8

8.1.2 Quantitation-2: Value-oriented, ratio

But in the ratio scale type we do not need a reference difference: a single reference object is enough (a meter has been defined in terms of a prototype meter bar). We can define the mean absolute error with a reference unit as:

$$MAE_{RU}(\sigma, \alpha) = \text{Avg}_{d \in \mathcal{D}} \left(\frac{|\alpha(d) - \sigma(d)|}{|\alpha(d_0)|} \right).$$

This can be applied to ratio scaled dimensions such as length or speed. Now we can prove the following theorem.

Theorem 11 (VOI_R, VOM_R and MAE) *The Mean Absolute Error with a reference unit is a value-oriented metric for the ratio scale type, i.e., MAE_{RU} satisfies VOI_R and VOM_R.*

8.1.3 Quantitation-3: Equivalence-oriented, interval

Let us consider the behaviour of an *equivalence-oriented metric for the interval scale type*. First it should be invariant under linear affine transformations of the system output. In addition, there must be a metric value increase if for every linear affine transformation for one assignment we can find a transformation for the other assignment which is value-closer to the gold. This is the case of the traditional Pearson correlation coefficient, which is defined as:

Table 10 The example described in the text

	α	σ_1	σ_2	σ_3	σ_4
d_1	Very positive	Very positive	Positive	Positive	Negative
d_2	Positive	Positive	Neutral	Negative	Negative

$$\text{CORR}(\sigma, \alpha) = \frac{\sum_i (\sigma(i) - \text{Avg}(\sigma))(\alpha(i) - \text{Avg}(\alpha))}{\sqrt{\sum_i (\sigma(i) - \text{Avg}(\sigma))^2} \sqrt{\sum_i (\alpha(i) - \text{Avg}(\alpha))^2}}$$

We can now prove the following theorem.

Theorem 12 (EOI_T, EOM_T and Pearson) *The Pearson correlation coefficient is an equivalence-oriented metric for the interval scale type, i.e., it satisfies EOI_T and EOM_T.*

8.1.4 Quantitation-4: Equivalence-oriented, ratio

Finally, we can also find equivalence oriented metrics at the ratio scale type. The most popular metric in this category is probably the cosine distance, defined as (where $\vec{\sigma} = \langle \sigma(i_1), \dots, \sigma(i_n) \rangle$ and $\vec{\alpha} = \langle \alpha(i_1), \dots, \alpha(i_n) \rangle$):

$$\text{COS}(\sigma, \alpha) = \frac{\vec{\sigma} \cdot \vec{\alpha}}{\|\vec{\sigma}\| \cdot \|\vec{\alpha}\|}$$

The strength of this similarity criterion is that it is not affected by proportionality transformations of assignments. According to the state of the art, the cosine distance is a good estimator for document similarity (in this case each item is the frequency of a word in the document). We can prove the following theorem.

Theorem 13 (EOI_R, EOM_R and cosine distance) *Whenever assignment values are positive, the cosine distance is an equivalence-oriented metric for the ratio scale type, i.e., it satisfies EOI_R and EOM_R.*

Table 9 shows the properties satisfied by these metrics according to the above theorems and the implication relationships between properties stated in Lemma 2 (the last column of the table is discussed in the following).

8.2 Ordinal classification: Value-oriented, ordinal

In Sect. 2.3 we highlighted that the Ordinal Classification task has not yet been analysed in depth, although several evaluation campaigns match this task and several authors have analysed the most popular metrics and have made some proposals (Gaudette and Japkowicz 2009; Baccianella et al. 2009; Cardoso and Sousa 2011). Our framework leads to this problem when considering value oriented metrics at ordinal scale, and provides two properties to be satisfied by metrics: VOI₀ and VOM₀. The first one states that the metric must be invariant under strict increasing functions (permissible transformation functions for the ordinal scale type) applied over both the system output and the gold: if we keep the relative order of gold and system

values, the metric must return the same result. VOM_0 states that approaching a value to the correct one must increase the system score.

Let us analyse the most popular metrics used in this task. On the one hand, the popular Accuracy metric is invariant at value-oriented nominal (VOI_N) and, therefore, it is also invariant at value-oriented ordinal scale, but it does not satisfy monotonicity (VOM_0). That is, it does not capture closeness to the gold value. An attempt in the literature to solve this gap is by means of *Accuracy with n* (Gaudette and Japkowicz 2009) which relaxes the range of values for a response to be accepted as matching. However, this solution does not solve the monotonicity problem for larger ordinal differences. Other authors proposed to use correlation coefficients, such as Pearson, Spearman, or Kendall; in particular, non parametric ones such as Kendall and Spearman are invariant at the ordinal scale, but they do not satisfy monotonicity, given that the maximum value of 1 can be achieved without returning the correct values. The Normalized Distance Performance Measure (NDPM) (Yao 1995) has the same behavior.

On the other hand, the Mean Average Error (MAE) and also the Mean Square Error (MSE) have been applied to this problem. They satisfy monotonicity (VOM_0) but at the cost of invariance, given that they take into account the interval distance between system and gold assigned values.

Let us focus on a specific example and consider two documents d_1 and d_2 , a gold α , and four system outputs $\sigma_1, \sigma_2, \sigma_3, \sigma_4$ with the values shown in Table 10. Notice that σ_1 resembles exactly the gold, σ_2 does not hit the target with the values, but it keeps the correct ordering as well as the correct distance between the categories (they are adjacent in this case), and σ_3 keeps the correct ordering but the second value moves further away from the correct result. Finally, σ_4 does not reflect the correct order of values and neither the correct value matching. Then, a metric should satisfy:

$$\mathcal{M}(\sigma_1, \alpha) > \mathcal{M}(\sigma_2, \alpha) > \mathcal{M}(\sigma_3, \alpha) > \mathcal{M}(\sigma_4, \alpha).$$

Accuracy is not able to discriminate among σ_2, σ_3 and σ_4 , given that all these system outputs fail in both target values. Therefore, value-oriented metrics for the nominal scale type are not adequate, since being closer to the real target should increase the score. Non parametric correlation coefficients (Kendall, Spearman, etc) do not discriminate among σ_1, σ_2 and σ_3 , given that all of them sort the documents in the correct way. The reason is that it is not a ranking problem either, so equivalence-oriented metric for the ordinal scale type are not appropriate. The linear correlation coefficient Pearson, an equivalence-oriented metric at interval scale, shows the same non discriminating effect as Spearman. In addition it also assumes that there exists the same interval between each category. However, we know that there are more categories between “Positive” and “Negative” than between “Positive” and “Neutral”, but we can not assert that the distance is the double, which is assumed by the Pearson coefficient. MAE and MSE would sort systems in a correct manner, given that they satisfy monotonicity (VOM_0). The limitation is that MAE and MSE are not invariant, and they require to assume an invariant value corresponding to each category, and thus pre-defined intervals between categories.

We claim that in this situation one must use value-oriented metrics for the ordinal scale type. The goal consists of reducing the distance between σ and α values, but at the same time, the distance can be defined only in ordinal terms. The more a prediction is far away from the target in ordinal terms, the more the system is penalized. The question is how to satisfy monotonicity and invariance simultaneously. We leave this open issue as future work.

9 Conclusions and future work

9.1 Summary

In this paper, we have defined a theoretical framework that explains the nature of evaluation metrics by grounding on measurement theory. Besides exploiting the traditional measurement theory, we have also introduced the concepts of value-oriented and equivalence-oriented closeness, for all scale types (nominal, ordinal, interval, ratio).

The theoretical results derived from the framework are:

- There is a clear correspondence between abstract tasks, metric kinds, and scale types. That is, classification, clustering, ranking, value prediction, and linear correlation oriented tasks can be interpreted as assignments for nominal, ordinal, interval, and ratio scale types, in which the closeness to the gold is evaluated at value or equivalence level.
- The definitions of value- and equivalence-oriented evaluation metrics match with the basic axioms stated in the literature for particular abstract tasks (strict monotonicity for classification, homogeneity and heterogeneity for clustering, and swapping for ranking): we only need to instantiate over the different scale types to infer these axioms.
- The proposed framework gives a single and unified explanation for most theoretical criticisms of classification, ranking, and clustering metrics found in the literature.
- The proposed framework explains the need for an interval and ratio units (i.e., meters, grades, etc.) when assignments are compared at the interval and ratio scale types.
- The proposed framework explains the popularity of Pearson coefficient and cosine distance when estimating the closeness of assignments at interval and ratio scale types.

Tables 8 and 9 summarize the aggregated analysis for all abstract tasks and metrics. Metrics used in different tasks match with the corresponding kind of metric according to our definition. The discarded metrics match with metric limitations actually identified in the literature. In addition, by filling the gap of value-oriented metrics for the ordinal scale type we understand how to evaluate tasks such as semantic textual similarity, recommendation, or polarity detection.

9.2 Practical consequences

We have already mentioned that our framework is not only theoretical. Let us summarize the practical contributions of this paper, mainly addressed to the communities of tasks and metric designers. First, the framework *gives a tool for selecting and checking the suitability of metrics*. One only needs to know: (i) in what scale the system output is defined, and (ii) if the goal consists of predicting values (i.e., classification, mean error, etc.) or relationships (i.e., clustering, ranking, or linear correlation). Second, the framework *helps users to distinguish between necessary properties and properties that depend on the particular characterization of the task*. The constraints found in the literature that match with our basic definitions of evaluation metric are strictly necessary for the corresponding abstract task. The other constraints depend on the particular task in which the evaluation is taking place. For instance, the priority constraint is a common desirable property to be satisfied by any ranking metric for the abstract task of ranking, while top-heaviness is task-dependent,

although necessary for IR. Third, the framework provides a tool for *defining metrics in situations in which, nowadays, the lack of suitable metrics enforces the use of several metrics*. The most clear example is the simultaneous use of Pearson and Spearman as well as error rate metrics when evaluating value prediction in an ordinal scale (i.e., sentiment polarity prediction).

9.3 Limits of this study and future developments

Our framework opens the door for evaluation metrics in empty theoretical spaces, such as value-oriented metrics for the ordinal scale type (that have not been proposed yet). However, it does not cover every scenario nor every metric property. The reason is that there exist particular axioms that depend on a particular task. One example is top-weightedness in the case of ranking. Another one is the ability of clustering metrics to avoid the combinatorial effect of element pairs in big clusters (counting pairs metrics fail on this). In classification, metrics can be grouped into classes depending of how random or non informative outputs are evaluated: this also depends on the particular task. However, we find this as a common situation in science. The first example that comes to mind is the exclusion of Euclid's fifth postulate to define different kinds of geometry.

Moreover, the framework is based on the assumption that system outputs and golds are assignments (of numerical values). This idea does not match with translation or summarization systems that generate text. Measurement theory has also a gap regarding the generation of structures. For instance, the inference of dependency trees from a text, or the construction of a hierarchical clustering do not fit into the idea of assignment. However, these tasks are often abstracted into classification or clustering abstract tasks.

In the literature one can find also composite metrics. For instance, diversity metrics in fact consider two golds simultaneously: the ordinal relevance of documents and their redundancy which is modeled as an assignment at the nominal scale type (information nuggets). These kinds of scenarios are not covered at the moment by the framework.

Finally, we have been focusing on the four classical scale types, but in measurement theory other scale types are proposed. Thus adding more tasks to our framework will be straightforward as long as they can be associated with a scale type. In this respect, we plan to further discuss the role of document filtering and its relations with other tasks.⁸

Acknowledgements This research has been partially supported by Grants *Vemodalen* (TIN2015-71785-R) and *MISMIS* (PGC2018-096212-B-C32) from the Spanish government, as well as by the Google Research Award *Axiometrics: Foundations of Evaluation Metrics in IR*.

⁸ Document filtering, i.e., discriminating relevant against irrelevant documents, is often considered as a task. The issue is probably more complex: filtering can be interpreted as: (i) a classification task (but with an implicit preference order, as one needs to know which is the positive/higher class), (ii) as a sort of “singularity” of ranking task with two levels only, for which some classification metrics are suitable, or (iii) a mixture of different tasks.

Appendix A: Measurement theory

In this appendix, we recall some basic concepts in measurement theory. Measurement theory is the discipline that studies the theoretical foundations of the activity of measuring. It is not a new field, and its roots are usually dated back to the seminal work by Stevens (1946). We try to make this paper self-contained: we therefore sacrifice some of the most technical issues when not necessary to our aims, as well as abuse the notation sometimes, and we provide some examples to help the intuition. A more formal and complete account can be found in the classical works by Suppes and Zinnes (1963), Roberts (1984), or Pedhazur and Schmelkin (1991).

Assignments and measurements

In measurement theory a measurement is defined as a function that assigns real numbers to elements of a set \mathcal{X} of objects or events in the real world. To be a measurement, the function must be a homomorphism, i.e., it must be such that the relationships in the so called *empirical relational structure*⁹ (i.e., the real world) are preserved in the *numerical relational structure* (i.e., the codomain of the function, the set of real numbers \mathbb{R}). For example, when measuring the weight of objects, we can have in the real world a relationship R that stands for “heavier than” and an operation \circ of “combination” (considering two objects together). A measurement will assign numbers such that R corresponds to “ $>$ ” and \circ to “ $+$ ”, thus defining a homomorphism from the empirical relational structure (the tuple $\langle \mathcal{X}, R, \circ \rangle$) to the numerical relational structure $\langle \mathbb{R}, >, + \rangle$.

For our purposes, we can start from a simpler situation: we define an assignment as a function that assigns a real number to a characteristic of an object or event.

Definition A.1 (*Assignment*) An assignment $\omega \in \Omega$ (we denote with Ω the set of all possible assignments) is a function that assigns values to objects in a set \mathcal{D} :

$$\omega : \mathcal{D} \longrightarrow \mathbb{R}.$$

Then, to be a measurement, an assignment must be a homomorphism (Roberts 1984; Suppes and Zinnes 1963; Pedhazur and Schmelkin 1991).

Definition A.2 (*Measurement*) Let $\langle \mathcal{D}, R_{\mathcal{D}} \rangle$ be an empirical relational structure and $\langle \mathbb{R}, R_{\mathbb{R}} \rangle$ be a numerical relational structure.¹⁰ A measurement $\psi \in \Psi$ (we denote with $\Psi \subset \Omega$ the set of all possible measurements) is an assignment of objects in a set \mathcal{D}

$$\psi : \mathcal{D} \longrightarrow \mathbb{R}$$

that is a homomorphism between the empirical relational structure and the numerical relational structure, i.e., $\forall x, y \in \mathcal{D} (R_{\mathcal{D}}(x, y) \implies R_{\mathbb{R}}(\psi(x), \psi(y)))$.

⁹ In this paper we prefer to use the term “structure” in place of the original “system” to avoid confusion, since the latter is already used for “IR system” and “system output”.

¹⁰ We assume that both empirical and numerical relational structures are pairs (a set with a single relationship). To be truly general, we should consider empirical and numerical relational structures having: (i) more than one relationship, (ii) relationships of different arity, and (iii) also functions besides relationships. This would mainly be a technical exercise and we refrain to do so to avoid unnecessary technical complications.

Table 11 For each scale type, the set of permissible transformation functions \mathcal{F}_T and the classic examples of Meaningful Statement s

T	\mathcal{F}_T (Permissible Transformation Functions)	$\widehat{\psi}_T^*$ (Meaningful Statements)
N	$\mathcal{F}_N = \{f \mid x = y \text{ iff } f(x) = f(y)\}$	$\widehat{\psi}_N^*(x, y) \iff x = y$
O	$\mathcal{F}_O = \{f \mid \text{if } x < y \text{ then } f(x) < f(y)\}$	$\widehat{\psi}_O^*(x, y) \iff x \geq y$
I	$\mathcal{F}_I = \{f \mid f(x) = a \cdot x + b, a > 0\}$	$\widehat{\psi}_I^*(x, y, z, w, k) \iff \frac{x-y}{z-w} = k$
R	$\mathcal{F}_R = \{f \mid f(x) = a \cdot x, a > 0\}$	$\widehat{\psi}_R^*(x, y, k) \iff \frac{x}{y} = k$

This definition is similar to the classical ones that can be found in the literature (Roberts 1984, pp. 51–52; Suppes and Zinnes 1963, pp. 4–8; Pedhazur and Schmelkin 1991, p. 17); it is very general and includes different kinds of measurements. Let us see some examples to emphasize this fact.

Example A.1 Let us consider three objects $\mathcal{D} = \{o_1, o_2, o_3\}$, and the measurement ψ_t of their temperature in Celsius degrees (we represent the function ψ_t as a set of pairs); then $\psi_t = \{\langle o_1, -5 \rangle, \langle o_2, 5 \rangle, \langle o_3, 15 \rangle\}$.

Example A.2 Let us consider three products $\mathcal{D} = \{p_1, p_2, p_3\}$, and the quality measurement ψ_q into high (we use the real number 2 for this), medium (1) or low (0); then $\psi_q = \{\langle p_1, 0 \rangle, \langle p_2, 1 \rangle, \langle p_3, 2 \rangle\}$.

Example A.3 Let us consider three students $\mathcal{D} = \{s_1, s_2, s_3\}$, and a measurement ψ_c which assigns each student to a class (foreign, local) identified with a number (1 = foreign, 2 = local): then $\psi_c = \{\langle s_1, 1 \rangle, \langle s_2, 2 \rangle, \langle s_3, 2 \rangle\}$.

It is important to remark that measurements are assignments (with further requirements). All the above $\psi_t, \psi_q,$ and ψ_c are of course assignments; to be measurements they need to be “correct”, in order to satisfy the homomorphism requirement. For example, the assignment $\psi'_t = \{\langle o_1, 5 \rangle, \langle o_2, -5 \rangle, \langle o_3, 15 \rangle\}$ would not be a measurement when using Celsius degrees, assuming that the values in ψ_t are indeed correct.

Given a measurement, one can of course make some inferences from it and derive some further properties such as equalities, greater than / less than relationships, differences, ratios, ratios of differences, etc. These properties have been named “numerical statements” (Suppes and Zinnes 1963) or simply “statements” (Roberts 1984). We use the latter term and we denote as ψ^* the set of statements that can be derived from the values assigned by ψ to the set of objects in \mathcal{D} . That is, $\psi^* = \{x \mid \psi \models x\}$. For instance, going back to our Example A.1, ψ_t^* contains statements like: $\forall i \in \{2, 3\}(\psi_t(o_i) > 0)$ (“temperatures of o_2 and o_3 are above zero”); $\psi_t(o_3) > \psi_t(o_2) > \psi_t(o_1)$; $\psi_t(o_3) = 3 \cdot \psi_t(o_2)$; and $\frac{\psi_t(o_3) - \psi_t(o_2)}{\psi_t(o_2) - \psi_t(o_1)} = 1$.

Let us remark that such statements can be either “value oriented” (i.e., $\psi_t(o_3) = 15$) or “relationship oriented” (i.e., $\psi_t(o_3) = 3 \cdot \psi_t(o_2)$ or $\psi_t(o_2) < \psi_t(o_3)$). This distinction is reconsidered and further discussed in the paper. Provided that there are no errors, all the statements derived from a measurement are numerically correct. For instance, in our Example A.1, $\psi_t(o_3) = -3 \cdot \psi_t(o_1)$ is numerically correct, while $\psi_t(o_3) = 4 \cdot \psi_t(o_1)$ is not. However, it makes no sense to assert that o_3 is minus three times hotter than o_2 ; whereas we can assert that there is the same difference of temperature between o_1, o_2 and o_3 . That is,

the statement $\frac{\psi_i(o_3) - \psi_i(o_2)}{\psi_i(o_2) - \psi_i(o_1)} = 1$ is numerically correct and makes sense, as well as the comparison $\psi_i(o_3) > \psi_i(o_2)$. In an analogous way, in the case of the Example A.2, it makes no sense to state that there is the same difference in quality between the products p_1, p_2 and p_3 although it is numerically true. We only know that some products have a higher quality than other products ($\psi_q(p_3) > \psi_q(p_2) > \psi_q(p_1)$). Finally, in the example of students and classes, there is no ordering relationship between classes: we can only say that two students belong to the same class or not.

According to the terminology used in measurement theory, we can summarize the above observations by saying that not all statements are *meaningful*. Meaningful statements are modeled in measurement theory on the basis of the concepts of *levels*, or *scales*.

Scales and scale types

The concepts of *scale* and *scale type* are central in measurement theory. The most typical categorization of scales is probably the one going from *Nominal* (N, the lowest scale type), through *Ordinal* (O) and *Interval* (I), to *Ratio* (R, the highest scale type). In our three examples above, temperature, product quality, and class assignment are I, O, and N scale types respectively. Other scales commonly used for temperature are of scale type I, like Fahrenheit, or R, like Kelvin. The choice of scale types is somehow arbitrary in the sense that there is not a single choice of which ones to define.

More formally, each scale type can be defined by means of an associated set of *permissible transformation functions*: each scale type corresponds to a particular set of functions.

Definition A.3 (*Permissible transformation functions*) The sets \mathcal{F}_T of permissible transformation functions for each scale type $T \in \{N, O, I, R\}$ are (see also Table 11):

- $\mathcal{F}_N = \{f \mid f \text{ is bijective, i.e., } x = y \text{ iff } f(x) = f(y)\};$
- $\mathcal{F}_O = \{f \mid f \text{ is strictly increasing monotonic, i.e., if } x < y \text{ then } f(x) < f(y)\};$
- $\mathcal{F}_I = \{f \mid f \text{ is a linear affinity, i.e., } f(x) = a \cdot x + b, a > 0\};$
- $\mathcal{F}_R = \{f \mid f \text{ is linear, i.e., } f(x) = a \cdot x, a > 0\}.$

We also write $f(\psi)$:

$$f(\psi) = \{\phi \mid \phi \in \Psi, \phi(x) = f(\psi(x))\}.$$

Given these permissible transformation functions, the potential measurements over a set of objects are grouped into equivalence classes according to the following equivalence relation.

Definition A.4 (*Measurement equivalence*) Two measurements ψ and ψ' are equivalent for the scale type T if there exists a permissible transformation function $f \in \mathcal{F}_T$ such that ψ and $f(\psi')$ are identical:

$$\psi \approx_T \psi' \iff \exists f \in \mathcal{F}_T (\psi = f(\psi')).$$

The set of all equivalent measurements is the equivalence class $[\psi]_T$.

The basic idea is that permissible transformation functions determine if two measurements are equivalent or not. For the nominal scale type, measurements which distribute objects into the same groups are equivalent, given that there exists a bijective function which transforms values from one measurement to the other. For instance, the student distribution in Example A.3 is equivalent to $\{\langle s_1, 4 \rangle, \langle s_2, 3 \rangle, \langle s_3, 3 \rangle\}$. The bijective function $f(x) = 5 - x$ transforms one measurement into another. For the ordinal scale type, measurements which order objects in the same way are equivalent. For instance, the measurement $\{\langle p_1, 4 \rangle, \langle p_2, 7 \rangle, \langle p_3, 10 \rangle\}$ is equivalent to the product quality measurement in Example A.2. The monotonic transformation function which confirms that the two measurement are equivalent is $f(x) = 4 + 3 \cdot x$.

These examples illustrate two characteristics of measurement scale types. The first one is that they are additional features of measurements. That is, two measurements defined on different scale types could assign the same numbers to objects, and therefore, the derived statements are identical, but the set of *meaningful* statements would not be the same.

The second characteristic is that there exists a natural subsumption across measurement scale types. For instance, equality relations are meaningful in every scale type. Greater than and less than make sense in every scale type except in \mathbb{N} . The difference between two values makes no sense in \mathbb{N} nor in \mathbb{O} . In other words, the more we consider “higher” scale types, the more we can state meaningful statements. That is, the permissible transformation functions for high measurement scale types (e.g., ratio, or interval) are also transformation functions for low measurement scale types (e.g., ordinal or nominal):

$$\mathcal{F}_R \subset \mathcal{F}_I \subset \mathcal{F}_O \subset \mathcal{F}_N. \tag{30}$$

Therefore, there exists also a subsumption between equivalence of measurements across scale types. That is, two equivalent measurements for a high scale type are also equivalent for lower scale types:

$$\psi \approx_R \psi' \implies \psi \approx_I \psi' \implies \psi \approx_O \psi' \implies \psi \approx_N \psi'.$$

We denote the ordering relationship on the set of the four scale types as

$$\mathbb{N} < \mathbb{O} < \mathbb{I} < \mathbb{R}, \tag{31}$$

and we speak of higher and lower scale types accordingly.

We briefly mention a property that is further discussed in the rest of the paper: there is a dependence between abstract tasks and scale types. Roughly, classification and clustering are based on the nominal scale type; ranking on the ordinal, and quantitation on the interval and ratio. However this dependence is not clear cut, as we discuss in more detail in the paper.

From assignment to measurements and back

Let us remark that although measurement theory is focused on measurements, some parts of it can be immediately generalized to assignments. More specifically, the notions of equivalent measurements and of equivalence class $[\psi]_T$ of a measurement ψ of Definition A.4 can be generalized to assignments in a straightforward way. Thus, given an assignment, it is possible to speak of an *equivalent assignment for a given scale type T*. The following definition formalizes the simple generalization of Definition A.4.

Definition A.5 (*Assignment equivalence*) Two assignments ω and ω' are equivalent for the scale type T if there exists a permissible transformation function $f \in \mathcal{F}_T$ such that ω and $f(\omega')$ are identical:

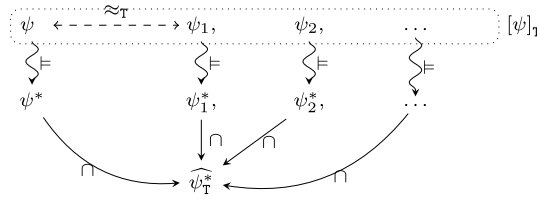


Fig. 2 Diagram of equivalent measurements (dashed line labelled with $\approx_{\mathcal{T}}$), equivalence classes (dotted rectangle labelled with $[\psi]_{\mathcal{T}}$), derivation (zigzag lines labelled with \models) of statements (ψ^*), and intersection (continuous lines labelled with \cap) to get the meaningful statements ($\psi_{\mathcal{T}}^*$)

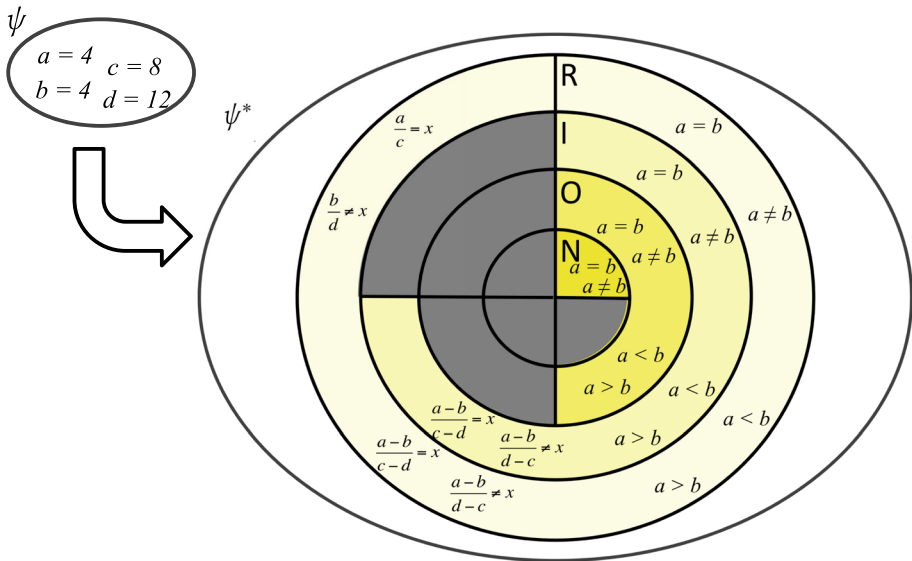


Fig. 3 (Meaningful) statements and scale types

$$\omega \approx_{\mathcal{T}} \omega' \iff \exists f \in \mathcal{F}_{\mathcal{T}}(\omega = f(\omega')).$$

The set of all equivalent assignments is the equivalence class $[\omega]_{\mathcal{T}}$.

The reason behind this definition is that if an assignment is not a measurement, then it means that there is no homomorphism between the empirical and numerical relational spaces, i.e., the assignment values are not correct w.r.t. the “real situation” (since measurement theory is interested in correct measurements, these cases are usually not taken into account); but in Definitions A.4 and A.5 there is no reference to the homomorphism.

Meaningfulness

We can now address the last element of measurement theory of interest here, namely the so called *meaningfulness problem* (Suppes and Zinnes 1963): as we have seen, given a

measurement ψ , not all the statements in ψ^* (sometimes not even some of those in ψ) are meaningful.

Commonly, measurement scale types are described in terms of what statements can be stated over two or more objects: the nominal scale type focusses on the equality relationships, the ordinal scale type incorporates the greater than and less than relationships between values, the interval scale type adds ratios of differences, and the ratio scale type includes ratios. Suppes and Zinnes (1963) define a “meaningful numerical statement” as a numerical statement which is constant under permissible transformation functions, that is, under an equivalent measurement. In this paper, we refer to this as *Meaningful Statement*.

Definition A.6 (Meaningful Statement) Given a measurement ψ of scale type T, the set of its Meaningful Statement s, denoted as $\widehat{\psi}_T^*$, contains the statements that are invariant across equivalent measurements:

$$\widehat{\psi}_T^* = \{x \in \psi^* \mid \forall \psi_i \approx_T \psi (x \in \psi_i^*)\} = \bigcap_{\forall \psi_i \approx_T \psi} \psi_i^*.$$

Figure 2 shows how *equivalence* and *meaningfulness* are related to each other. At the top left of the figure, there is the measurement ψ . On its right there are its equivalent measurements (ψ_i) on the same domain and for the same scale type T. All of them belong to the equivalence class $[\psi]_T$. For instance, they could be different measurement of the ordinal scale type which keep the same order across object values. Each measurement produces different derived statements (ψ^*, ψ_i^*), but the meaningful statements for the scale type T are those that are shared by every derived measurement.

According to this definition, we can infer certain meaningful statements for each scale type. Figure 3 shows a graphical representation of some meaningful statements for each scale type. As the figure shows, there is a subsumption across scale types, being the equality the most basic relationship in the nominal scale type. It is important to remark that those listed in the figure are not all the possible meaningful statements. For example, one could compute the arithmetic mean of some temperatures in Example A.1 as $\text{mean}(\psi_i(o_1), \psi_i(o_2)) = 0$, $\text{mean}(\psi_i(o_2), \psi_i(o_3)) = 10$ and state that the difference between the two means is the same as $\psi_i(o_2) - \psi_i(o_1)$, and this would be true under any permissible transformation function in \mathcal{F}_T , i.e., meaningful.

The most important aspect of this figure is the grey area that highlights how some statements for certain scale types are not meaningful. For instance, the ordinal relationship between values ($a > b$) is not invariant across equivalent measurements of the nominal scale type. Therefore it is hidden by the grey area.

As shown in the last column of Table 11, equality is a boolean 2-ary meaningful relationships for the \mathbb{N} scale type; greater than (\geq) is a meaning relationship for the \mathbb{O} scale type; the ratio of differences is a 5-ary meaningful relationship for the \mathbb{I} scale type; and the ratio is a 3-ary meaningful relationship for the \mathbb{R} scale type.

It is also important to remark that, perhaps surprisingly but obviously, for any measurement ψ , the assignments of single values in ψ (e.g., $\psi(d_1) = 6$) are *never* meaningful:¹¹ none of them is invariant under the corresponding permissible transformation functions. In other terms, to find something meaningful one needs to look inside the set of derived

¹¹ The only case for which this would be different is the sometimes proposed *absolute* scale type, for which the only permissible transformation function is identity.

numerical statements ψ^* . For example, $\psi_r(d_1) = -5$ of Example A.1 is false when changing to Fahrenheit, although “the real temperatures are always the same”. This statement captures something that is not “in the empirical relational structure” (Suppes and Zinnes 1963), and does not represent properties of the “true” temperatures; rather, it is an artefact of a particular measurement approach.

It is also important to understand that meaningfulness is not always an important, or even desired, property. If one is given the task “measure temperature using the Celsius scale”, then providing an equivalent measurement in Fahrenheit is not a satisfactory achievement of the task. Also this issue is discussed in the paper.

Appendix B: Proofs

Proof of Lemma 1 We need to prove that a value $x \in \mathbb{R}$ is (non-strictly) closer to a reference r than y for each scale type T respectively if and only if the conditions in Table 5 are satisfied.

According to Definition 2, a value $x \in \mathbb{R}$ is (non-strictly) closer to a reference r than y for a certain scale type T if there exists a permissible transformation function in \mathcal{F}_T such that it is (non-strictly) closer. We prove each scale type independently.

Let us consider the nominal case and non-strict closeness. By contraposition:

$$\neg(r \neq y \vee r = x) \iff \neg(r \neq y) \wedge \neg(r = x) \iff r = y \wedge r \neq x.$$

Therefore, for any nominal transformation function applied over r, x and y , given that they preserve the equalities and inequalities:

$$\begin{aligned} \forall f \in \mathcal{F}_N(f(r) = f(y) \wedge f(r) \neq f(x)) \\ \implies \forall f \in \mathcal{F}_N(|f(r) - f(x)| > 0 \wedge |f(r) - f(y)| = 0) \\ \implies \forall f \in \mathcal{F}_N(|f(r) - f(x)| > |f(r) - f(y)|) \\ \implies \neg \exists f \in \mathcal{F}_N(|f(r) - f(x)| \leq |f(r) - f(y)|) \\ \implies \neg \exists f \in \mathcal{F}_N(f(x) \leq^{f(r)} f(y)) \implies \neg(x \leq_N^r y) \end{aligned}$$

and therefore

$$x \leq_N^r y \implies (r \neq y \vee r = x).$$

Let us consider the other direction:

$$(r \neq y \vee r = x) \implies x \leq_N^r y.$$

First, note that:

$$(r \neq y \vee r = x) \iff (r \neq y \vee r = x = y).$$

If $r \neq y$ we can find a permissible transformation function such that $f(r) \geq f(x) > f(y)$. If $r = x = y$, then the identity transformation also satisfies $f(r) \geq f(x) \geq f(y)$. Therefore, $x \leq_N^r y$.

The strict case can be inferred starting from the definition in Formula (4) and by exploiting the non-strict case:

$$\begin{aligned}
 x \prec_{\mathbb{N}}^r y &\iff (x \preccurlyeq_{\mathbb{N}}^r y) \wedge \neg(y \preccurlyeq_{\mathbb{N}}^r x) \\
 &\iff (r \neq y \vee r = x) \wedge \neg(r \neq x \vee r = y) \\
 &\iff (r \neq y \vee r = x) \wedge r = x \wedge r \neq y \\
 &\iff (r = x \wedge r \neq y).
 \end{aligned}$$

Let us turn to the ordinal case. We have to prove that:

$$x \preccurlyeq_0^r y \iff \neg(r \geq y > x \vee x > y \geq r).$$

We start by observing that:

$$\begin{aligned}
 \neg(r \geq y > x \vee x > y \geq r) &\iff \neg(r \geq y > x) \wedge \neg(x > y \geq r) \\
 &\iff (\neg(r \geq y) \vee \neg(y > x)) \wedge (\neg(x > y) \vee \neg(y \geq r)) \\
 &\iff (r < y \vee y \leq x) \wedge (x \leq y \vee y < r) \\
 &\iff (r < y \wedge (x \leq y \vee y < r)) \vee (y \leq x \wedge (x \leq y \vee y < r)) \\
 &\iff (r < y \wedge x \leq y) \vee (r < y \wedge y < r) \vee (y \leq x \wedge x \leq y) \vee (y \leq x \wedge y < r) \\
 &\iff (r < y \wedge x \leq y) \vee (x = y) \vee (y \leq x \wedge y < r).
 \end{aligned}$$

In all these three cases, we can define a permissible transformation function for the ordinal scale type such that:

$$|f(r) - f(x)| \leq |f(r) - f(y)|.$$

Therefore:

$$x \preccurlyeq_0^r y \iff \neg(r \geq y > x \vee x > y \geq r).$$

On the other hand, if $\neg(\neg(r \geq y > x \vee x > y \geq r))$, then $r \geq y > x$ or $x > y \geq r$. Given that in both cases y is strictly closer to r than x , we can not find a permissible ordinal transformation function such that $|f(r) - f(x)| < |f(r) - f(y)|$, and therefore $\neg(x \preccurlyeq_0^r y)$. Then, by contraposition:

$$x \preccurlyeq_0^r y \implies \neg(r \geq y > x \vee x > y \geq r).$$

The strict case is easily derived, again starting from (4) and by exploiting the non-strict case:

$$\begin{aligned}
 x \prec_0^r y &\iff x \preccurlyeq_0^r y \wedge \neg(y \preccurlyeq_0^r x) \\
 &\iff (r \geq x > y \vee y > x \geq r) \wedge \neg(r \geq y > x \vee x > y \geq r) \\
 &\iff (r \geq x > y \vee y > x \geq r) \wedge (r < y \vee y \leq x) \wedge (x \leq y \vee y < r) \\
 &\iff (r \geq x > y) \vee (y > x \geq r).
 \end{aligned}$$

Finally, regarding the scale types I and R, we want to prove that:

$$x \preccurlyeq_{I,R}^r y \iff |r - x| \leq |r - y|.$$

We can prove that the condition on the right hand side is invariant under the permissible transformation functions for the ratio scale type:

Table 12 The nominal permissible transformation function

x	1	2	3	6	5	4	10	20	30
$f(x)$	6	5	4	1	2	3	10	20	30

$$|r - x| \leq |r - y| \iff a \cdot |r - x| \leq a \cdot |r - y| \iff |a \cdot r - a \cdot x| \leq |a \cdot r - a \cdot y|.$$

Therefore, the condition is valid for the R scale type. In addition, we know that adding a constant does not affect the absolute difference ($|r - x| = |r + b - (x + b)|$). Therefore the condition is invariant also for the interval scale type:

$$|r - x| \leq |r - y| \iff |a \cdot r + b - (a \cdot x + b)| \leq |a \cdot r + b - (a \cdot y + b)|.$$

Therefore

$$x \leq'_{I,R} y \implies \exists f \in \mathcal{F}_{I,R} (|f(r) - f(x)| \leq |f(r) - f(y)|) \implies |r - x| \leq |r - y|$$

and

$$|r - x| \leq |r - y| \implies \exists f \in \mathcal{F}_{I,R} (|f(r) - f(x)| \leq |f(r) - f(y)|) \implies x \leq'_{I,R} y.$$

The proof for the strict case is immediate. □

Proof of Lemma 2 We prove each item in the lemma independently, with the exception of items (c) and (d) that are proved together.

- (a) VOI_T states the property in Formula (13) for any $f \in \mathcal{F}_T$. Obviously, the same property holds for f in any subset of \mathcal{F}_T , and if $T < T'$ then $\mathcal{F}_{T'} \subset \mathcal{F}_T$ (see Formulas (30) and (31)). In other terms, if a metric value does not change by applying any transformation function of a certain scale type, then the metric is also invariant for transformation functions of higher scale types: satisfying VOI for a certain scale type implies satisfying VOI for higher scale types.
- (b) The proof for EOI is analogous to that for VOI: see the previous case (a) and start from Formula (15).
- (c) This item is proved together with the next one.
- (d) The proof is divided into two parts. First, we focus on the nominal scale type and we prove that VOI_N is incompatible with both VOM (item (c) of the lemma) and EOM (item (d) of the lemma) at higher scale types. Let us assume that VOI_N holds. Then, by reduction to absurdity, we prove that if VOM or EOM hold, then we find a contradictory implication and therefore, they are incompatible. Let us consider the following particular assignments for $\mathcal{D} = \{d_1, d_2, d_3\}$:

$$\begin{aligned} \sigma &= (1, 2, 3) \\ \sigma' &= (6, 5, 4) \\ \alpha &= (10, 20, 30). \end{aligned}$$

Note that if we apply the nominal permissible transformation function $x \mapsto f(x)$ defined in Table 12, then $\forall d \in \mathcal{D} \ f(\sigma(d)) = \sigma'(d)$, $f(\sigma'(d)) = \sigma(d)$, and $f(\alpha(d)) = \alpha(d)$. Now, if a metric satisfies VOI_N , this implies that the metric must be

Table 13 The ordinal permissible transformation function

x	1	2	3	10	100	1000	10000	20000	30000
$f(x)$	10	100	1000	2000	3000	4000	10000	20000	30000

invariant under any nominal transformation $f \in \mathcal{F}_N$ applied over both the system output and the gold. Therefore:

$$\mathcal{M}(\sigma, \alpha) = \mathcal{M}(f(\sigma), f(\alpha)) = \mathcal{M}(\sigma', \alpha) \tag{32}$$

(when clear from the context we omit the subscript on the metric). On the other hand,

$$\forall d \in \mathcal{D} (\sigma(d) < \sigma'(d) < \alpha(d)). \tag{33}$$

Therefore, according to the closeness definition at ordinal, interval, and ratio scale types, for any $d \in \mathcal{D}$, $\sigma'(d)$ is closer to $\alpha(d)$ than $\sigma(d)$, i.e., $\forall T \in \{0, I, R\} (\sigma' \triangleleft_T^\alpha \sigma)$. Now, by reduction to absurdity, if we assume that VOM at 0, I, or R scale types holds, we can derive that

$$\mathcal{M}(\sigma', \alpha) > \mathcal{M}(\sigma, \alpha) \tag{34}$$

(see Formula (14)), which contradicts (32). Therefore VOM can not be satisfied at higher scale types than nominal. This proves item (c).

Now, let us prove that EOM can not be satisfied at higher than nominal scale types. We can assert that $\sigma'(d_1) > \sigma'(d_2) > \sigma'(d_3)$ and $\alpha(d_1) < \alpha(d_2) < \alpha(d_3)$. Therefore, σ' and α are not equivalent at 0, I, or R scale types. This implies that the assignment α is strictly equivalence-closer to itself than σ' to α . On the other hand, the ratio, interval, and ordinal permissible transformation function $f(x) = 10 * x$ makes $f(\sigma(i)) = \alpha(i)$.

Therefore, for any transformation $f'(\sigma')$ at 0, I, or R, there exists a transformation $f(\sigma) = \alpha$ that is value-closer to α than $f'(\sigma')$, whereas the converse does not hold: according to Formulas (7) and (8), $\forall T \in \{0, I, R\} (\sigma \sqsubset_T^\alpha \sigma')$. Therefore, according to EOM at 0, I, or R (see Formula (16)),

$$\mathcal{M}(\sigma, \alpha) > \mathcal{M}(\sigma', \alpha), \tag{35}$$

which again contradicts (32): also EOM can not be satisfied at higher scale types than nominal. Now we turn to the second part of the proof, which has a similar structure: we focus on the ordinal scale type and we prove, again by reduction to absurdity, that VOI_0 is incompatible with both VOM and EOM at higher scale types. Let us consider the following particular assignments:

$$\begin{aligned} \sigma &= (1, 2, 3) \\ \sigma' &= (10, 100, 1000) \\ \alpha &= (10000, 20000, 30000). \end{aligned}$$

Note that if we apply the ordinal permissible transformation function $x \mapsto f(x)$ in Table 13, then $f(\sigma) = \sigma'$ and $f(\alpha) = \alpha$. Now, satisfying VOI_0 implies that the metric must be invariant under any ordinal permissible transformation function $f \in \mathcal{F}_0$ applied over both the system output and the gold, i.e., Formula (32) must hold for any such f . On the other hand, (33) still holds. Therefore, according to the closeness definition at interval and ratio scale types, for any $d \in \mathcal{D}$, $\sigma'(d)$ is closer to $\alpha(d)$ than

$\sigma(d)$, i.e., $\forall T \in \{I, R\}(\sigma' \triangleleft_T^\alpha \sigma)$. Now, if by reduction to absurdity we assume that VOM at I or R scale types holds, we can derive (34) (see Formula (14)), which contradicts (32): VOM can not be satisfied at higher scale types than ordinal. In addition, the interval and ordinal permissible transformation function $f(x) = 10000 * x$ makes $f(\sigma) = \alpha$. And given that σ' and α are not linearly correlated, since

$$\sigma'(i) = 10^{(\alpha(i)/10000)},$$

we can assert that σ' and α are not equivalent at I or R scale types. Therefore, for any transformation $f'(\sigma')$ at I or R, there exists a transformation $f(\sigma) = \alpha$ that is value-closer to α than $f'(\sigma')$, whereas the converse does not hold: according to Formulas (7) and (8), $\forall T \in \{I, R\}(\sigma \sqsubset_T^\alpha \sigma')$. Therefore, according to EOM at I or R (see Formula (16)) we have that (35) holds, which again contradicts (32): also EOM can not be satisfied at higher scale types than 0.

- (e) The proof is again by reduction to absurdity. Let us consider the following assignments:

$$\begin{aligned} \sigma &= (1, 2, 3) \\ \sigma' &= (10, 10, 30) \\ \alpha &= (10, 20, 30). \end{aligned}$$

$f(x) = 10 * x$ is a permissible transformation function for every scale type. Given that $\alpha(i) = \sigma(i) * 10$, we can say that σ and α are equivalent at every scale type, whereas there does not exist any bijective relationship between σ' and α , i.e., there does not exist any nominal permissible transformation function between σ' and α . Therefore, there does not exist any permissible transformation at any scale type. It follows that, for any transformation at any scale type of σ' , the transformation $\sigma(i) * 10$ is value-closer to α . Thus, we can say that for every scale type, σ is equivalence-closer to α than σ' . Then, if EOM is satisfied by \mathcal{M} at any scale type, we obtain:

$$\mathcal{M}(\sigma, \alpha) > \mathcal{M}(\sigma', \alpha). \tag{36}$$

On the other hand, σ' is strict value-closer to α than σ at nominal scale type, given that $\sigma(d_1) \neq \sigma'(d_1) = \alpha(d_1)$, $\sigma(d_2) \neq \sigma'(d_2) \neq \alpha(d_2)$ and $\sigma(d_3) \neq \sigma'(d_3) = \alpha(d_3)$. Then, by subsumption of scale types, we can say that σ' is strict value-closer than σ at every scale type. Therefore, if \mathcal{M} satisfies VOM at any scale type then:

$$\mathcal{M}(\sigma, \alpha) < \mathcal{M}(\sigma', \alpha),$$

which contradicts the EOM implication (36).

- (f) It is enough to consider that, according to Lemma 1, the conditions for closeness and strict closeness at interval and ratio scale types are equivalent.
- (g) The proof is by reduction to absurdity. Let us consider the following assignments:

$$\begin{aligned} \sigma &= (1, 2, 3) \\ \sigma' &= (10, 20, 30) \\ \alpha &= (10, 20, 30). \end{aligned}$$

We note that two equal values are strictly closer than two different values for every scale type. Therefore, $\sigma' \triangleleft_T^\alpha \sigma$ for every scale type T. Therefore, according to VOM_T:

$$\mathcal{M}(\sigma', \alpha) < \mathcal{M}(\sigma, \alpha).$$

On the other hand, σ and σ' are equivalent at every scale type, given that the ratio permissible function $f(x) = 10 * x$ states $f(\sigma(i)) = \sigma'(i)$. By subsumption across scale types, σ and σ' are equivalent at every scale type. Therefore, according to EOI_T at every scale type:

$$\mathcal{M}(\sigma', \alpha) = \mathcal{M}(\sigma, \alpha),$$

which contradicts the previous result.

- (h) First, we prove that in general, two equivalent assignments are strictly equivalence-closer than two non equivalent assignments:

$$\sigma' \in [\alpha]_T \wedge \sigma \notin [\alpha]_T \implies \sigma' \sqsubset_T^\alpha \sigma.$$

This is straightforward, given that we can find a permissible transformation function from σ' to α but not from σ to α . Therefore, the transformation of σ' is strictly value-closer than any transformation of σ .

Let us now consider the following assignments:

$$\begin{aligned} \sigma_1 &= (100, 10, 1) \\ \sigma_2 &= (1, 10, 100) \\ \sigma_3 &= (11, 21, 31) \\ \sigma' &= (10, 20, 30) \\ \alpha &= (10, 20, 30). \end{aligned}$$

We have that:

$$\begin{aligned} \forall T (\sigma' \in [\alpha]_T) & \quad \text{and} \\ \forall T \geq R(\sigma_3 \notin [\alpha]_T) & \quad \text{and} \quad \forall T \leq I(\sigma_3 \in [\alpha]_T) \\ \forall T \geq I(\sigma_2 \notin [\alpha]_T) & \quad \text{and} \quad \forall T \leq O(\sigma_2 \in [\alpha]_T) \\ \forall T \geq O(\sigma_1 \notin [\alpha]_T) & \quad \text{and} \quad \forall T \leq N(\sigma_1 \in [\alpha]_T). \end{aligned}$$

Now, by reduction to absurdity, we can derive the following contradictions:

$$\begin{aligned} \text{EOM}_R &\implies \mathcal{M}(\sigma', \alpha) > \mathcal{M}(\sigma_3, \alpha) \text{ and } \forall T \leq I(\text{EOI}_T \implies \mathcal{M}(\sigma', \alpha) = \mathcal{M}(\sigma_3, \alpha)) \\ \text{EOM}_T &\implies \mathcal{M}(\sigma', \alpha) > \mathcal{M}(\sigma_2, \alpha) \text{ and } \forall T \leq O(\text{EOI}_T \implies \mathcal{M}(\sigma', \alpha) = \mathcal{M}(\sigma_2, \alpha)) \\ \text{EOM}_0 &\implies \mathcal{M}(\sigma', \alpha) > \mathcal{M}(\sigma_1, \alpha) \text{ and } \forall T \leq N(\text{EOI}_T \implies \mathcal{M}(\sigma', \alpha) = \mathcal{M}(\sigma_1, \alpha)). \end{aligned}$$

□

Proof of Corollary 1 The proofs are simply based on the well known property that $\neg(A \wedge B)$ is the same as $A \implies \neg B$: by applying this to Lemma 2 (c), (d), (e) (g), and (h), it is immediate to obtain Formulas (17), (19), (21), (23), and (25). Conversely, (18), (20), (22), (24), and (26) are simply derived respectively from (17), (19), (21), (23), and (25) by contraposition (i.e., $A \implies \neg B$ is the same as $B \implies \neg A$). □

Proof of Corollary 2 The corollary statement is the same as saying that satisfying both EOM and EOI at a certain scale type, or both VOM and VOI at certain scale type, is not compatible with any other combination, with the only exception of the case of VOM and

VOI at ratio and interval scale types. Let us prove it by analyzing all the possible combinations. First, according to Lemma 2 (g), VOM_T and $EOI_{T'}$ are incompatible for any pair of scale types T and T' .

Therefore, value- and equivalence-oriented metrics are always incompatible definitions. Then, let us prove the incompatibility of metrics within the same family:

- Value-oriented metrics at nominal and ordinal scale types. According to Lemma 2 (c), satisfying VOI implies that VOM can not be satisfied at higher scale types, satisfying VOM implies that VOI can not be satisfied at lower scale types.
- Value-oriented metrics at interval or ratio scale types. According to Lemma 2 (c), satisfying VOM at interval or ratio scale types implies that VOI can not be satisfied at nominal and ordinal scale types.
- Equivalence-oriented metrics. According to Lemma 2 (h), satisfying EOI implies that EOM can not be satisfied at lower scale types (or, satisfying EOM implies that EOI can not be satisfied at higher scale types).

□

Proof of Theorem 1 The GMON axiom (Axiom 1) states that:

$$\forall d \in \mathcal{D}(\alpha(d) = \sigma(d) \vee \alpha(d) \neq \sigma'(d)) \wedge (\exists d \in \mathcal{D}(\alpha(d) = \sigma(d) \neq \sigma'(d))) \\ \implies \mathcal{M}(\sigma, \alpha) > \mathcal{M}(\sigma', \alpha).$$

By comparing it with the VOM_N axiom (Formula (14)) we see that the consequent is the same, so we just need to prove that the two antecedents are equivalent, i.e., that:

$$\sigma \prec_N^\alpha \sigma' \iff \forall d \in \mathcal{D}(\alpha(d) = \sigma(d) \vee \alpha(d) \neq \sigma'(d)) \wedge (\exists d \in \mathcal{D}(\alpha(d) = \sigma(d) \neq \sigma'(d))). \tag{37}$$

By applying (6) and Lemma 1 for the nominal scale, we have that:

$$\sigma \prec_N^\alpha \sigma' \iff \forall d \in \mathcal{D}(\sigma(d) \preceq_N^{\alpha(d)} \sigma'(d)) \wedge \exists d \in \mathcal{D}(\sigma(d) \prec_N^{\alpha(d)} \sigma'(d)) \\ \iff \forall d \in \mathcal{D}(\alpha(d) \neq \sigma'(d) \vee \alpha(d) = \sigma(d) = \sigma'(d)) \wedge \\ \exists d \in \mathcal{D}(\alpha(d) = \sigma(d) \wedge \alpha(d) \neq \sigma'(d)),$$

that is equivalent to the right hand side of (37). □

Proof of Theorem 2 We want to prove that the EOM axiom and the Generalized Homogeneity/Completeness axiom (GHC, Axiom 2) are equivalent. As in the previous proof of Theorem 1, this is the same as proving that the conditions for a metric score increase in the two axioms are equivalent. The conditions for the GHC axiom are Formulas (27) and (28); the condition for the EOM axiom is the antecedent in (16).

Part A

First, we prove that the conditions for the GHC axiom implies the condition for the EOM axiom. Using (8), the condition in EOM axiom can be rewritten as $\sigma \sqsubseteq_N^\alpha \sigma' \wedge \neg(\sigma' \sqsubseteq_N^\alpha \sigma)$. In

Part A.1 we prove that GHC axiom conditions implies $\neg(\sigma' \sqsubseteq_N^\alpha \sigma)$. In Part A.2 we prove that GHC conditions implies $\sigma \sqsubseteq_N^\alpha \sigma'$.

Part A.1.

First, we can state the following equivalences:

$$\begin{aligned}
 \neg(\sigma' \sqsubseteq_N^\alpha \sigma) &\iff \neg(\forall \sigma_e \in [\sigma]_N (\exists \sigma'_e \in [\sigma']_N (\sigma'_e \leq_N^\alpha \sigma_e))) \\
 &\iff \exists \sigma_e \in [\sigma]_N \neg(\exists \sigma'_e \in [\sigma']_N (\sigma'_e \leq_N^\alpha \sigma_e)) \\
 &\iff \exists \sigma_e \in [\sigma]_N (\forall \sigma'_e \in [\sigma']_N \neg(\sigma'_e \leq_N^\alpha \sigma_e)) \\
 &\iff \exists \sigma_e \in [\sigma]_N \left(\forall \sigma'_e \in [\sigma']_N \neg(\forall d \in \mathcal{D} (\sigma'_e(d) \leq_N^{\alpha(d)} \sigma_e(d))) \right) \\
 &\iff \exists \sigma_e \in [\sigma]_N \left(\forall \sigma'_e \in [\sigma']_N (\exists d \in \mathcal{D} (\sigma_e(d) <_N^{\alpha(d)} \sigma'_e(d))) \right).
 \end{aligned}
 \tag{38}$$

(where we have used (7), (5), and (4)).

Now, looking at Formula (28) in the GHC conditions, we can observe that it is divided into two or-ed terms. If the first term $(\sigma'(d_1) = \sigma'(d_2) \wedge \alpha(d_1) \neq \alpha(d_2) \wedge \sigma(d_1) \neq \sigma(d_2))$ is true, we can define an assignment σ_e such that

$$\sigma_e(d_1) = \alpha(d_1) \neq \sigma_e(d_2) = \alpha(d_2).$$

Given that $\sigma'(d_1) = \sigma'(d_2)$, these relationships will be preserved by any equivalent assignment σ'_e . Therefore, for at least one of the two documents d_1 or d_2 , σ_e will be strictly closer to α than σ'_e .

If the second term $(\sigma'(d_1) \neq \sigma'(d_2) \wedge \alpha(d_1) = \alpha(d_2) \wedge \sigma(d_1) = \sigma(d_2))$ is true, we can define an assignment $\sigma_e \in [\sigma]$ such that:

$$\sigma_e(d_1) = \alpha(d_1) = \sigma_e(d_2) = \alpha(d_2).$$

Given that $\sigma'(d_1) \neq \sigma'(d_2)$, again, these relationships will be preserved by any equivalent assignment σ'_e . Therefore, for at least one of the two documents d_1 or d_2 , σ_e will be strictly closer to α than σ'_e .

Therefore, in both cases, the assertion:

$$\exists \sigma_e \in [\sigma]_N \left(\forall \sigma'_e \in [\sigma']_N \left(\exists d \in \mathcal{D} \left(\sigma_e(d) <_T^{\alpha(d)} \sigma'_e(d) \right) \right) \right)$$

is true, and therefore, using (38), $\neg(\sigma' \sqsubseteq_N^\alpha \sigma)$.

Part A.2.

Now we prove that the GHC conditions imply $\sigma \sqsubseteq_N^\alpha \sigma'$, which according to (7) can be expressed as:

$$\sigma \sqsubseteq_N^\alpha \sigma' \iff \forall \sigma'_e \in [\sigma']_N (\exists \sigma_e \in [\sigma]_N (\sigma_e \leq_N^\alpha \sigma'_e)).$$

That is, we have to prove that for any assignment σ'_e in $[\sigma']$ there exists an assignment σ_e in $[\sigma]$ such that it is value-closer to α than σ'_e for the scale type N. To do so, we make the following four steps: we define an assignment σ_e which depends on the α , σ , and σ' values (A.2.1); we prove that it is an assignment, i.e., it is a function that assigns a value to each

item (A.2.2); we prove that σ_e belongs to $[\sigma]$ (A.2.3); and finally we prove that σ_e is value-closer to α than σ' (A.2.4).

Part A.2.1. Let us define the following assignment:

$$\sigma_e(d) = \begin{cases} \alpha(d_j) & \text{if } \exists d_j(\sigma'_e(d_j) = \alpha(d_j) \wedge \sigma(d_j) = \sigma(d)) \\ \sigma(d) + K & \text{otherwise.} \end{cases} \tag{39}$$

This assignment assigns the correct α value to documents that are correctly assigned by σ'_e , and extends this assignment to other documents, namely all those that are assigned the same σ value as d and are correctly assigned by σ'_e , for preserving the consistency with σ , that is, in order to belong to $[\sigma]$. In other terms, $\sigma_e(d)$ assigns the correct value α when σ'_e does. In addition, σ_e is consistent (equivalent) with σ at nominal level, given that the equality relationships are preserved in both α values and $\sigma(d) + K$ values. The K value is just a number large enough to avoid overlapping with α values (for example, if α and σ assign values in $\{0, 1\}$, we can define $K = 2$).

Part A.2.2. We have to prove that σ_e is an assignment. It is enough to ensure that two values are not assigned simultaneously by $\sigma_e(d)$. For this, we will prove that if there exist two documents d_j in Formula (39) that are assigned a value by $\sigma'_e(d)$, then these values are the same one. That is, if

$$\exists d_j(\sigma'_e(d_j) = \alpha(d_j) \wedge \sigma(d_j) = \sigma(d))$$

and

$$\exists d_k(\sigma'_e(d_k) = \alpha(d_k) \wedge \sigma(d_k) = \sigma(d))$$

then $\alpha(d_j) = \alpha(d_k)$. That is, we have to prove that

$$\sigma(d_k) = \sigma(d_j) \wedge \sigma'_e(d_j) = \alpha(d_j) \wedge \sigma'_e(d_k) = \alpha(d_k) \implies \alpha(d_j) = \alpha(d_k).$$

Here we use the GHC Axiom. According to the second part of Formula (27)

$$(\sigma'(d_j) \neq \sigma'(d_k) \wedge \alpha(d_j) \neq \alpha(d_k)) \implies \sigma(d_j) \neq \sigma(d_k),$$

therefore:

$$\sigma(d_j) = \sigma(d_k) \implies (\sigma'(d_j) = \sigma'(d_k) \vee \alpha(d_j) = \alpha(d_k)).$$

Then:

$$\begin{aligned} \sigma(d_k) = \sigma(d_j) \wedge \sigma'_e(d_j) = \alpha(d_j) \wedge \sigma'_e(d_k) = \alpha(d_k) &\implies \\ (\sigma'(d_j) = \sigma'(d_k) \vee \alpha(d_j) = \alpha(d_k)) \wedge \sigma'_e(d_j) = \alpha(d_j) \wedge \sigma'_e(d_k) = \alpha(d_k). \end{aligned}$$

Given that σ'_e is equivalent to σ' , this implies that:

$$\begin{aligned} (\sigma'_e(d_j) = \sigma'_e(d_k) \vee \alpha(d_j) = \alpha(d_k)) \wedge \sigma'_e(d_j) = \alpha(d_j) \wedge \sigma'_e(d_k) = \alpha(d_k) \\ \implies \alpha(d_j) = \alpha(d_k). \end{aligned}$$

Therefore, two values are not assigned simultaneously by $\sigma_e(d)$ and it is an assignment.

Part A.2.3. Now, we have to prove the assignment σ_e belongs to the class $[\sigma]$. We know that at the nominal scale type, two assignments σ and σ_e are equivalent if:

$$\sigma(d_i) = \sigma(d_j) \iff \sigma_e(d_i) = \sigma_e(d_j).$$

Suppose that $\sigma(d_i) = \sigma(d_j)$. Then, if

$$\exists d_k (\sigma'_e(d_k) = \alpha(d_k) \wedge \sigma(d_k) = \sigma(d_i))$$

then

$$\exists d_k (\sigma'_e(d_k) = \alpha(d_k) \wedge \sigma(d_k) = \sigma(d_j))$$

and therefore (see (39)) $\sigma_e(d_i) = \sigma_e(d_j) = \alpha(d_k)$. In the other cases:

$$\sigma_e(d_i) = \sigma_e(d_j) = \sigma(d_j) + K.$$

On the other hand, if $\sigma(d_i) \neq \sigma(d_j)$, then according to (39) the only case in which $\sigma_e(d_i) = \sigma_e(d_j)$ is if

there exist two documents d_k and d_l such that

$$\exists d_k (\sigma'_e(d_k) = \alpha(d_k) \wedge \sigma(d_k) = \sigma(d_i))$$

$$\exists d_l (\sigma'_e(d_l) = \alpha(d_l) \wedge \sigma(d_l) = \sigma(d_j)),$$

and $\sigma_e(d_i) = \sigma_e(d_j) = \alpha(d_k) = \alpha(d_l)$. Then, in this case:

$$\sigma'_e(d_k) = \sigma'_e(d_l) \wedge \alpha(d_k) = \alpha(d_l) \wedge \sigma(d_i) \neq \sigma(d_j).$$

Given that $\sigma'_e \in [\sigma']$:

$$\sigma'(d_k) = \sigma'(d_l) \wedge \alpha(d_k) = \alpha(d_l) \wedge \sigma(d_i) \neq \sigma(d_j),$$

which is not compatible with the GHC conditions.

Therefore, the proposed assignment σ_e belongs to the class $[\sigma]$.

Part A.2.4. Finally, we need to prove that σ_e is value-closer to α than σ'_e for the scale type \mathbb{N} :

$$\sigma'_e(d) = \alpha(d) \implies \sigma_e(d) = \alpha(d).$$

This is solved by definition of σ_e .

Therefore, according to Lemma 1 we can conclude that for every document d , $\sigma_e(d)$ is non strictly closer to $\alpha(d)$ than $\sigma'_e(d)$:

$$\forall d \in \mathcal{D} (\sigma_e(d) \preceq^{\alpha(d)} \sigma'_e(d)).$$

Therefore, according to Definition 3, σ_e is non-strictly value-closer to α than σ'_e for the nominal scale type:

$$\sigma_e \preceq_{\mathbb{N}}^{\alpha} \sigma'_e.$$

Given that this is true for every assignment $\sigma'_e \in [\sigma']_{\mathbb{N}}$, we can conclude that σ is non-strictly equivalence-closer to α than σ' for the nominal scale type:

$$\sigma \sqsubseteq_{\mathbb{N}}^{\alpha} \sigma'.$$

So, the EOM condition is satisfied.

Part B

Now, we will prove that the EOM condition is enough for the GHC axiom conditions. The GHC conditions include Formulas (27) and (28). We will prove that they are satisfied in parts B.1 and B.2 respectively.

Part B.1.

The EOM condition states that, for any assignment σ'_e in $[\sigma']_{\mathbb{N}}$ we can find an assignment σ_e in $[\sigma]_{\mathbb{N}}$ such that $\sigma_e \leq_{\mathbb{N}}^{\alpha} \sigma'_e$. Let us consider the two conditions in Formula (27). Then, if

$$\sigma'(d_i) = \sigma'(d_j) \wedge \alpha(d_i) = \alpha(d_j)$$

there exists an assignment σ'_e in the equivalence class such that:

$$\sigma'_e(d_i) = \sigma'_e(d_j) = \alpha(d_i) = \alpha(d_j).$$

Given that there must exist an assignment in the class $[\sigma]_{\mathbb{N}}$ closer to α , then the assignments in $[\sigma]_{\mathbb{N}}$, including σ , must satisfy:

$$\sigma(d_i) = \sigma(d_j).$$

Therefore, the first condition in Formula (27) in the GHC axiom is satisfied:

$$\sigma'(d_i) = \sigma'(d_j) \wedge \alpha(d_i) = \alpha(d_j) \implies \sigma(d_i) = \sigma(d_j).$$

We can apply the same reasoning for the situation

$$\sigma'(d_i) \neq \sigma'(d_j) \wedge \alpha(d_i) \neq \alpha(d_j).$$

That is, there exists an assignment σ'_e in the equivalence class $[\sigma']_{\mathbb{N}}$ such that:

$$\sigma'_e(d_i) = \alpha(d_i) \neq \sigma'_e(d_j) = \alpha(d_j).$$

Given that there must exist an assignment in the class $[\sigma]_{\mathbb{N}}$ closer to α , then the assignments in $[\sigma]_{\mathbb{N}}$, including σ , must satisfy:

$$\sigma(d_i) \neq \sigma(d_j).$$

Therefore, the second condition in Formula (27) in the GHC axiom is also satisfied:

$$\sigma'(d_i) \neq \sigma'(d_j) \wedge \alpha(d_i) \neq \alpha(d_j) \implies \sigma(d_i) \neq \sigma(d_j).$$

Part B.2.

In addition, the conditions in Formula (28) are also necessary. Otherwise, if they were false then:

$$\begin{aligned}
 & \forall d_i, d_j (\neg(\sigma'(d_i) = \sigma'(d_j) \wedge \alpha(d_i) \neq \alpha(d_j) \wedge \sigma(d_i) \neq \sigma(d_j))) \\
 & \quad \wedge \forall d_i, d_j (\neg(\sigma'(d_i) \neq \sigma'(d_j) \wedge \alpha(d_i) = \alpha(d_j) \wedge \sigma(d_i) = \sigma(d_j))) \\
 \iff & \forall d_i, d_j ((\sigma'(d_i) \neq \sigma'(d_j) \vee \alpha(d_i) = \alpha(d_j) \vee \sigma(d_i) = \sigma(d_j)) \\
 & \quad \wedge ((\sigma'(d_i) = \sigma'(d_j) \vee \alpha(d_i) \neq \alpha(d_j) \vee \sigma(d_i) \neq \sigma(d_j))) \\
 \iff & (\sigma(d_i) \neq \sigma(d_j) \wedge \alpha(d_i) \neq \alpha(d_j) \implies \sigma'(d_i) \neq \sigma'(d_j)) \\
 & \quad \wedge (\sigma(d_i) = \sigma(d_j) \wedge \alpha(d_i) = \alpha(d_j) \implies \sigma'(d_i) = \sigma'(d_j)).
 \end{aligned}$$

Therefore, being σ_e an assignment belonging to the class $[\sigma]_{\mathbb{N}}$, we can find an assignment σ'_e in $[\sigma']_{\mathbb{N}}$ such that:

$$\sigma_e(d_i) = \alpha(d_i) \implies \sigma'_e(d_i) = \alpha(d_i).$$

This assignment is:

$$\sigma'_e(d) = \begin{cases} \alpha(d) & \text{if } \sigma_e(d) = \alpha(d) \\ \sigma'(d) + K & \text{otherwise.} \end{cases}$$

Therefore, according to Lemma 1, we can conclude that σ'_e is closer to α than σ_e :

$$\sigma'_e \sqsubseteq_{\mathbb{N}}^{\alpha} \sigma_e.$$

Given that this is true for any assignment σ_e in the class $[\sigma]_{\mathbb{N}}$, we can conclude that:

$$\sigma' \sqsubseteq_{\mathbb{N}}^{\alpha} \sigma$$

which contradicts

$$\sigma \sqsubset_{\mathbb{N}}^{\alpha} \sigma'.$$

Therefore, EOM and GHC axioms are equivalent. □

Proof of Theorem 3 We want to prove that EOM at ordinal scale type is equivalent to PRI, the Priority Axiom. For this, it is enough to prove that the EOM and PRI conditions are equivalent. On the one hand, Priority Axiom (Axiom 3) states that if two contiguous documents are swapped in concordance with the gold then the metric score increases. In other terms, if¹²

$$\text{rank}_{\sigma'}(d_1) = \text{rank}_{\sigma'}(d_2) + 1 \wedge \alpha(d_1) > \alpha(d_2)$$

and being σ the results of swapping d_1 and d_2 in σ' then $\mathcal{M}(\sigma, \alpha) > \mathcal{M}(\sigma', \alpha)$.

Given that $\mathcal{M}(\sigma, \alpha)$ is a function which depends exclusively on the σ and α assignment values, swapping documents in the ranking σ with equal value in α does not affect $\mathcal{M}(\sigma, \alpha)$. Therefore, by transitivity, we can say that the condition of the Priority Axiom (Formula 29) is equivalent to $S(\sigma, \sigma')$, which denotes the fact that σ can be obtained by swapping documents in the σ' ranking without contradicting α .

¹² Note that by increasing rank ($\text{rank}_{\sigma}(d) < \text{rank}_{\sigma}(d')$) we decrease relevance ($\sigma(d) > \sigma(d')$); see Footnote 12

On the other hand, the EOM condition (the antecedent in (16)) holds when σ is equivalence-closer than σ' . That is, for all ordinal transformation of σ' there exists a transformation of σ strictly closer to α . We will denote this condition as C_{EOM} . Then, we need to prove that:

$$C_{\text{EOM}} \iff S(\sigma, \sigma').$$

Part A

First, let us prove that:

$$C_{\text{EOM}} \implies S(\sigma, \sigma')$$

which, by contraposition, is equivalent to:

$$\neg S(\sigma, \sigma') \implies \neg C_{\text{EOM}}. \quad (40)$$

If $\sigma = \sigma'$ then (40) is true, given that neither $S(\sigma, \sigma')$ nor C_{EOM} hold. In the other case, whenever $\sigma \neq \sigma'$, if there does not exist any (correct) swapping sequence from σ' to σ , then we can assert that there exist at least two documents which are correctly sorted in σ' but not in σ . That is:

$$\sigma(d'_1) > \sigma(d'_2) \wedge \alpha(d_1) > \alpha(d_2) \wedge \sigma(d_1) < \sigma(d_2).$$

Under this condition, let us consider the permissible transformation function $f_{\sigma'}$ for the ordinal scale type such that

$$f_{\sigma'}(\sigma'(d_1)) = \alpha(d_1) \text{ and } f_{\sigma'}(\sigma'(d_2)) = \alpha(d_2).$$

Now, by reduction to absurdity, let us assume that the EOM condition C_{EOM} holds, i.e., that there exists another monotonic function f_{σ} such that it is value-closer to α for every d . Therefore:

$$f_{\sigma}(\sigma(d_1)) = \alpha(d_1) \wedge f_{\sigma}(\sigma(d_2)) = \alpha(d_2).$$

Since f_{σ} is a monotonic function, we have:

$$f_{\sigma}(\sigma(d_1)) < f_{\sigma}(\sigma(d_2)).$$

But this contradicts $\alpha(d_1) > \alpha(d_2)$. Therefore, we can not find a transformation f_{σ} of σ such that all values are closer to α than the $f_{\sigma'}$ transformation of σ' . Therefore, the condition of EOM can not be satisfied and this proves (40).

Part B

Now, let us prove that the conditions for PRI (Formula (29)) imply satisfying the condition of the EOM axiom. The priority axiom states that if the rankings σ and σ' are equivalent except in the case of d_1 and d_2 where:

$$\alpha(d_1) > \alpha(d_2) \wedge \sigma(d_1) > \sigma(d_2) \wedge \sigma'(d_1) < \sigma'(d_2), \quad (41)$$

then σ must achieve a higher score than σ' .

Let σ'_e be any assignment in the equivalence class $[\sigma']$. Then, we consider an assignment σ_e such that:

$$\forall d \neq d_2 (\sigma_e(d) = \sigma'_e(d)) \wedge \sigma_e(d_2) = \sigma'_e(d_1) - k \tag{42}$$

where k is a number small enough such that:

$$\sigma(d) < \sigma(d_2) \implies \sigma_e(d) < \sigma_e(d_2).$$

The intuition is that σ_e moves d_2 down in the rank as much as it is needed to correctly (w.r.t. α) swap it with d_1 , without affecting the other relationships. Then, we can assert that σ_e belongs to the equivalence class $[\sigma]$, given that: (i) for every pair of documents different than d_1 and d_2 , σ , σ' , σ_e and σ'_e keep the same ordinal relationships; (ii) the relationship between d_1 and d_2 is the same in σ (according to Formula (41)) and in σ_e (according to Formula (42)); and (iii) given that d_1 and d_2 are contiguous in σ' , σ'_e , σ and σ_e , the relationships with the rest of the documents are the same in all assignments.

Now, let us consider an assignment α' in $[\alpha]$ such that $\alpha'(d_1) = \sigma'_e(d_1) = \sigma_e(d_1)$. Then we can assert that σ_e is value-closer to α' than σ'_e (i.e., $\sigma_e \sqsubseteq_1^{\alpha'} \sigma'_e$): for every $d \neq d_2$, σ_e is closer given that

$$\forall d \neq d_2 (\sigma_e(d) = \sigma'_e(d)),$$

and regarding d_2 , being k small enough and knowing that $\alpha'(d_1) > \alpha'(d_2)$, we have that

$$\sigma'_e(d_2) > \sigma'_e(d_1) - k = \sigma_e(d_2) = \alpha'(d_1) - k > \alpha'(d_2).$$

That is, for any transformation σ'_e of σ' we can find a transformation σ_e of σ which is value-closer to α' . Therefore:

$$\sigma \sqsubseteq_0^{\alpha'} \sigma'. \tag{43}$$

As the last step of the proof, we need to prove that equivalence-closeness is invariant under equivalent assignments. That is:

$$\forall \alpha' \in [\alpha] (\sigma_1 \sqsubseteq_0^{\alpha'} \sigma_2 \iff \sigma_1 \sqsubseteq_0^{\alpha} \sigma_2). \tag{44}$$

To prove this, it is enough to consider that equivalence-closeness at ordinal scale type is invariant across ordinal transformations of σ_1 or σ_2 :

$$\sigma_1 \sqsubseteq_0^{\alpha'} \sigma_2 \iff f_1(\sigma_1) \sqsubseteq_0^{\alpha'} f_2(\sigma_2)$$

and that value-closeness at ordinal scale type is invariant when applying the same transformation to the three assignments:

$$\sigma_1 \sqsubseteq_0^{\alpha'} \sigma_2 \iff f(\sigma_1) \sqsubseteq_0^{f(\alpha')} f(\sigma_2).$$

Therefore, being f_α the transformation such that $f_\alpha(\alpha') = \alpha$:

$$\sigma_1 \sqsubseteq_0^{\alpha'} \sigma_2 \iff f_\alpha(\sigma_1) \sqsubseteq_0^{f_\alpha(\alpha')} f_\alpha(\sigma_2) \iff f_\alpha(\sigma_1) \sqsubseteq_0^{\alpha} f_\alpha(\sigma_2) \iff \sigma_1 \sqsubseteq_0^{\alpha} \sigma_2.$$

In our case, from (43) we can thus derive that:

$$\sigma \sqsubseteq_0^\alpha \sigma'.$$

Finally, for a transformation σ_e in $[\sigma]$ such that $\sigma_e(d_1) = \alpha(d_1)$ and $\sigma_e(d_2) = \alpha(d_2)$ we can not find a transformation of σ'_e in $[\sigma']$ closer to α than σ_e due to $\sigma'(d_1) < \sigma'(d_2)$. Therefore, we can assert that the priority axiom conditions implies:

$$\sigma \sqsubset_0^\alpha \sigma'.$$

That is, the EOM conditions for the ordinal scale type. □

Proof of Theorem 4 The proof is straightforward. For the nominal scale type, applying a permissible transformation function (bijective function) is equivalent to changing the names of classes. Given that the contingency matrix is based on equalities and inequalities between system output and gold, modifying the class names for both the system output and the gold does not affect the matrix. □

Proof of Theorem 5 It is enough to prove that these metrics satisfy the Generalized Monotonicity Axiom (GMON), which is equivalent to VOM at nominal scale according to Theorem 1. GMON states that:

$$\forall d \in \mathcal{D}(\alpha(d) = \sigma(d) \vee \alpha(d) \neq \sigma'(d)) \wedge \exists d \in \mathcal{D}(\alpha(d) = \sigma(d) \neq \sigma'(d)) \\ \implies \mathcal{M}(\sigma, \alpha) > \mathcal{M}(\sigma', \alpha).$$

For Accuracy and Macro-Average Accuracy the proof is straightforward. The metric Accuracy directly counts how many documents are classified correctly. Therefore, the first assignment σ will achieve always a higher accuracy than σ' under the previous conditions. The Macro-Average Accuracy consists of averaging Accuracy scores across classes, that is, values in the gold. It can be expressed as:

$$\text{MAAC} = \text{Avg}_v \left(\frac{\text{card}(\{d \mid \alpha(d) = v \wedge \sigma(d) = v\})}{\text{card}(\{d \mid \alpha(d) = v\})} \right).$$

Under the GMON conditions, the assignment σ achieves the same Accuracy as σ' over all values, excepting one of them in which its Accuracy is higher.

The Phi Correlation is a metric that can be applied only over two classes, that is, two values in both assignments. In terms of the contingency matrix, it is expressed as (TP, TN, FP, FN, stand for True Positive, True Negative, False Positive, and False Negative, respectively):

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

Under the GMON conditions, there are two possibilities. TP is increased at the cost of FN or TN is increased at the cost of FP. In the first case:

$$\frac{(TP + 1) \cdot TN - FP \cdot (FN - 1)}{\sqrt{(TP + 1 + FN - 1) \cdot (TN + FP) \cdot (TP + 1 + FP) \cdot (TN + FN - 1)}}.$$

Now, if we derive the function:

$$\frac{(TP + x) \cdot TN - FP \cdot (FN - x)}{\sqrt{(TP + x + FN - x) \cdot (TN + FP) \cdot (TP + x + FP) \cdot (TN + FN - x)}}$$

with respect to x , we obtain a function which is positive for every x value. We can apply exactly the same procedure to the second case. Therefore, the function is an increasing monotonic function. That is, the *Phi* value increases under the GMON conditions. \square

Proof of Theorem 6 The proof is straightforward. It is enough to say that the information given by a partition is invariant across permissible transformation functions for the nominal scale type, i.e., the bijective functions (applied to one partition or both). Therefore, any function that takes a partition as input is invariant and it satisfies EOI. \square

Proof of Theorem 7 We will prove that Entropy satisfies GHC (Axiom 2), and therefore, according to Theorem 2, it also satisfies EOM at the nominal scale type. We can state that the GHC conditions (see Formulas (27) and (28)) concern two clustering outputs σ and σ' such that the only differences are either: (i) σ splits clusters of σ' into clusters containing items from different classes in the gold α , or (ii) σ joins clusters of σ' containing items in the same class in α . In other terms, in case (ii), if σ' breaks the relationship between items from the same class in α , then σ' is producing a pair error; in case (i), if σ' joins clusters containing items from different classes in α , σ' is producing incorrect relationships. In both cases the metric value for σ' must be lower than σ .

Knowing this, let us prove now that joining clusters containing items from the same class or splitting clusters in this way increases Entropy. Let us focus on the joining case (ii) first. Let $c_1, c_2 \in \mathcal{V}(\sigma')$ be the two clusters containing elements from the same class $l \in \mathcal{V}(\alpha)$ in the gold, and that are joined in σ . The Entropy of both clusters (i.e., for σ') is:

$$E(c_1, \alpha) = - \sum_{l \in \mathcal{V}(\alpha)} P(\alpha(i) = l \mid \sigma'(i) = c_1) \cdot \log P(\alpha(i) = l \mid \sigma'(i) = c_1) = 0$$

$$E(c_2, \alpha) = - \sum_{l \in \mathcal{V}(\alpha)} P(\alpha(i) = l \mid \sigma'(i) = c_2) \cdot \log P(\alpha(i) = l \mid \sigma'(i) = c_2) = 0$$

(we slightly abuse the notation introduced in Sect. 7.2 by replacing assignments with clusters), and Entropy after the join (i.e., for σ) is zero as well:

$$E(c, \alpha) = - \sum_{l \in \mathcal{V}(\alpha)} P(\alpha(i) = l \mid \sigma(i) = c) \cdot \log P(\alpha(i) = l \mid \sigma(i) = c) = 0$$

(where $c = c_1 \cup c_2$). That is, joining these clusters does not affect Entropy. However, Class Entropy for the label l is (before the joining process, i.e., for σ'):

$$\begin{aligned}
 CE(\sigma', l) &= - \sum_{c \in \mathcal{V}(\sigma')} P(\sigma'(i) = c \mid \alpha(i) = l) \cdot \log P(\sigma'(i) = c \mid \alpha(i) = l) \\
 &= - P(\sigma'(i) = c_1 \mid \alpha(i) = l) \cdot \log P(\sigma'(i) = c_1 \mid \alpha(i) = l) \\
 &\quad - P(\sigma'(i) = c_2 \mid \alpha(i) = l) \cdot \log P(\sigma'(i) = c_2 \mid \alpha(i) = l) \\
 &\quad - \sum_{c \in \mathcal{V}(\sigma') \setminus \{c_1, c_2\}} P(\sigma'(i) = c \mid \alpha(i) = l) \cdot \log P(\sigma'(i) = c \mid \alpha(i) = l).
 \end{aligned}
 \tag{45}$$

Now let us remark that, for any $a, b \in (0..1)$ and $a + b < 1$, we have

$$0 > \log(a + b) > \log(a) \text{ and } 0 > \log(a + b) > \log(b),$$

and, given that a and b are positive:

$$a \cdot \log(a) + b \cdot \log(b) < a \cdot \log(a + b) + b \cdot \log(a + b) = (a + b) \cdot \log(a + b).$$

Therefore:

$$-a \cdot \log(a) - b \cdot \log(b) > -(a + b) \cdot \log(a + b).$$

By exploiting this property in (45) we obtain:

$$\begin{aligned} CE(\sigma', l) &> - (P(\sigma'(i) = c_1 \mid \alpha(i) = l) + P(\sigma'(i) = c_2 \mid \alpha(i) = l)) \\ &\cdot \log (P(\sigma'(i) = c_1 \mid \alpha(i) = l) + P(\sigma'(i) = c_2 \mid \alpha(i) = l)) \\ &- \sum_{c \in \mathcal{V}(\sigma') \setminus \{c_1, c_2\}} P(\sigma'(i) = c \mid \alpha(i) = l) \cdot \log P(\sigma'(i) = c \mid \alpha(i) = l) \end{aligned}$$

which corresponds with the entropy for the class l after the joining process (i.e., for σ). Therefore, Class Entropy decreases when joining pure clusters: being the evaluation score inversely correlated with the entropy values, this proves that the metric value for σ' is lower than σ .

Now let us consider the case (i) in which one cluster in σ' is split into two clusters $c_1, c_2 \in \mathcal{V}(\sigma)$. When splitting two clusters $c_1, c_2 \in \mathcal{V}(\sigma)$ containing items from different classes in the gold:

$$\begin{aligned} E(c_1 \cup c_2, \alpha) &= - \sum_{l \in \mathcal{V}(\alpha)} P(\alpha(i) = l \mid \sigma(i) = c_1) \cdot \log P(\alpha(i) = l \mid \sigma(i) = c_1) \\ &- \sum_{l \in \mathcal{V}(\alpha)} P(\alpha(i) = l \mid \sigma(i) = c_2) \cdot \log P(\alpha(i) = l \mid \sigma(i) = c_2) \\ &= E(c_1, \alpha) + E(c_2, \alpha). \end{aligned}$$

But Entropy is averaged by considering the size of clusters. Therefore, given that $E(c_1, \alpha)$ and $E(c_2, \alpha)$ are positive numbers ($-P(x) \log(x) \in (0.. + \infty)$):

$$\begin{aligned} &|c_1| \cdot E(c_1, \alpha) + |c_2| \cdot E(c_2, \alpha) \\ &< |c_1| \cdot (E(c_1, \alpha) + E(c_2, \alpha)) + |c_2| \cdot (E(c_1, \alpha) + E(c_2, \alpha)) \\ &= (|c_1| + |c_2|) \cdot (E(c_1, \alpha) + E(c_2, \alpha)), \end{aligned}$$

we can say that Entropy decreases when splitting clusters with items belonging to different classes (from σ' to σ). In addition, Class Entropy is not affected by this splitting, given that the distribution of classes across clusters is not affected.

Therefore, remembering that Entropy and Class Entropy are inversely correlated with the evaluation metric, they satisfy GHC and thus EOM at the nominal scale type. \square

Proof of Theorem 8 The proof is straightforward. In our framework, system outputs are value assignments, that is, document relevance scores estimated by systems. Therefore, by definition, the resulting ranking is invariant under ordinal transformation functions of these scores. In other words, applying any permissible transformation function (strict monotonic function) over the system output does not affect the resulting ranking. That is, evaluation

metrics for ranking are invariant across equivalent assignments for the ordinal scale type. □

Proof of Theorem 9 For this proof, it is enough to say that (Amigó et al. 2013), proved that these metrics satisfy PRI, which is equivalent to EOM at ordinal scale according to Theorem 3. □

Proof of Theorem 10 We need to prove that the Mean Absolute Error with a reference difference is a value-oriented metric for the interval scale type. VOI is satisfied, given that the metric is invariant across linear transformations of α and σ values, as it is immediately seen by considering linear affine transformations like $(X' = aX + b)$ and observing that:

$$\frac{|X' - Y'|}{|Z' - W'|} = \frac{|aX + b - aY - b|}{|aZ + b - aW - b|} = \frac{|aX - aY|}{|aZ - aW|} = \frac{|a(X - Y)|}{|a(Z - W)|} = \frac{|X - Y|}{|Z - W|}.$$

VOM is satisfied, given that by applying (6) and Lemma 1 for the interval scale type, we have that

$$\sigma \triangleleft_{\tau}^{\alpha} \sigma' \iff \forall d \in \mathcal{D}(\sigma(d) \leq_{\tau}^{\alpha(d)} \sigma'(d)) \wedge \exists d \in \mathcal{D}(\sigma(d) <_{\tau}^{\alpha(d)} \sigma'(d)),$$

and this is equivalent to

$$\forall d \in \mathcal{D}(|\sigma(d) - \alpha(d)| \leq |\sigma'(d) - \alpha(d)|) \wedge \exists d \in \mathcal{D}(|\sigma(d) - \alpha(d)| < |\sigma'(d) - \alpha(d)|).$$

Therefore:

$$\text{Avg}_{d \in \mathcal{D}} \left(\frac{|\alpha(d) - \sigma(d)|}{|\alpha(d_0) - \alpha(d'_0)|} \right) < \text{Avg}_{d \in \mathcal{D}} \left(\frac{|\alpha(d) - \sigma'(d)|}{|\alpha(d_0) - \alpha(d'_0)|} \right).$$

□

Proof of Theorem 11 VOI is satisfied, given that the metric is invariant across linear transformation of α and σ values, as it is immediately seen by considering linear transformations like $(X' = aX)$ and observing that:

$$\frac{|X' - Y'|}{|Z'|} = \frac{|aX - aY|}{|aZ|} = \frac{|X - Y|}{|Z|}.$$

The proof for VOM is analogous to that of previous Theorem 10:

$$\text{If } \sigma \triangleleft_{\tau}^{\alpha} \sigma' \text{ then } \text{Avg}_{d \in \mathcal{D}} \left(\frac{|\alpha(d) - \sigma(d)|}{|\alpha(d_0)|} \right) < \text{Avg}_{d \in \mathcal{D}} \left(\frac{|\alpha(d) - \sigma'(d)|}{|\alpha(d_0)|} \right).$$

□

Proof of Theorem 12 We want to prove that the Pearson coefficient is an equivalence-oriented metric for the interval scale type, i.e., it satisfies EOI_{τ} and EOM_{τ} . A first remark is that the invariance under interval permissible transformation functions (i.e., linear functions $y = a \cdot x + b$) is a well known property of Pearson linear correlation coefficient. Therefore, EOI_{τ} is satisfied.

Then, we need to prove that Pearson satisfies EOM at interval scale type. That is, Pearson correlation with a gold α increases when going from an assignment σ_1 to an equivalence-closer to α assignment σ_2 :

$$\sigma_2 \sqsubset_1^\alpha \sigma_1 \implies \text{CORR}(\sigma_2, \alpha) > \text{CORR}(\sigma_1, \alpha). \tag{46}$$

We prove it by contraposition; therefore, we need to prove:

$$\text{CORR}(\sigma_2, \alpha) \leq \text{CORR}(\sigma_1, \alpha) \implies \neg(\sigma_2 \sqsubset_1^\alpha \sigma_1), \tag{47}$$

i.e., that if $\text{CORR}(\sigma_2, \alpha) \leq \text{CORR}(\sigma_1, \alpha)$ then σ_2 is not equivalence-closer to α than σ_1 .

We start by considering the normalization $\bar{\sigma}$ of an assignment σ

$$\bar{\sigma}(i) = \frac{\sigma(i) - \text{Avg}(\sigma)}{\text{Dev}(\sigma)}$$

(where $\text{Avg}(\sigma)$ represents the average value and we use $\text{Dev}(\sigma)$ for the standard deviation, in order to avoid confusion with the notation of σ as an assignment) that has the following properties:

$$\sum_{i \in \mathcal{D}} \bar{\sigma}(i) = 0, \tag{48}$$

$$\sum_{i \in \mathcal{D}} \bar{\sigma}(i)^2 = 1, \tag{49}$$

and

$$\text{CORR}(\bar{\sigma}, \bar{\alpha}) = \sum_{i \in \mathcal{D}} \bar{\sigma}(i)\bar{\alpha}(i). \tag{50}$$

Given that Pearson is invariant under interval permissible transformations, it holds that:

$$\text{CORR}(\sigma_2, \alpha) \leq \text{CORR}(\sigma_1, \alpha) \iff \text{CORR}(\bar{\sigma}_2, \bar{\alpha}) \leq \text{CORR}(\bar{\sigma}_1, \bar{\alpha}),$$

Then, according to (50)

$$\begin{aligned} \text{CORR}(\bar{\sigma}_2, \bar{\alpha}) \leq \text{CORR}(\bar{\sigma}_1, \bar{\alpha}) &\iff \sum_{i \in \mathcal{D}} \bar{\sigma}_2(i)\bar{\alpha}(i) \leq \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i)\bar{\alpha}(i) \\ &\iff -2 \sum_{i \in \mathcal{D}} \bar{\sigma}_2(i)\bar{\alpha}(i) \geq -2 \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i)\bar{\alpha}(i). \end{aligned} \tag{51}$$

Now, let us consider any linear transformation σ'_2 of the normalized assignment $\bar{\sigma}_2$. That is: $\sigma'_2(i) = a \cdot \bar{\sigma}_2(i) + b$ with $a > 0$. Then:

$$\begin{aligned} \sum_{i \in \mathcal{D}} (\sigma'_2(i) - \bar{\alpha}(i))^2 &= \sum_{i \in \mathcal{D}} (a\bar{\sigma}_2(i) + b - \bar{\alpha}(i))^2 \\ &= \sum_{i \in \mathcal{D}} (a^2\bar{\sigma}_2(i)^2 + (b - \bar{\alpha}(i))^2 + 2a\bar{\sigma}_2(i)(b - \bar{\alpha}(i))) \\ &= a^2 \sum_{i \in \mathcal{D}} \bar{\sigma}_2(i)^2 + \sum_{i \in \mathcal{D}} (b - \bar{\alpha}(i))^2 + 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_2(i)b - 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_2(i)\bar{\alpha}(i). \end{aligned}$$

According to (49):

$$= a^2 + \sum_{i \in \mathcal{D}} (b - \bar{\alpha}(i))^2 + 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_2(i)b - 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_2(i)\bar{\alpha}(i).$$

Then, according to (51) and given that $a > 0$:

$$\geq a^2 + \sum_{i \in \mathcal{D}} (b - \bar{\alpha}(i))^2 + 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_2(i)b - 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i)\bar{\alpha}(i).$$

Finally, applying (48) twice to replace a zero term with another zero term:

$$\begin{aligned} &= a^2 + \sum_{i \in \mathcal{D}} (b - \bar{\alpha}(i))^2 + 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i)b - 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i)\bar{\alpha}(i) \\ &= a^2 + \sum_{i \in \mathcal{D}} (b - \bar{\alpha}(i))^2 + 2ab \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i) - 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i)\bar{\alpha}(i) \\ &= a^2 \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i)^2 + \sum_{i \in \mathcal{D}} (b - \bar{\alpha}(i))^2 - 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i)\bar{\alpha}(i) \\ &= \sum_{i \in \mathcal{D}} (a\bar{\sigma}_1(i) + b - \bar{\alpha}(i))^2. \end{aligned}$$

That is, reading the equality/inequality chain, we can assert that for every value $a > 0$ and b :

$$\begin{aligned} \text{CORR}(\sigma_2, \alpha) \leq \text{CORR}(\sigma_1, \alpha) &\implies \\ \sum_{i \in \mathcal{D}} (a\bar{\sigma}_2(i) + b - \bar{\alpha}(i))^2 &\geq \sum_{i \in \mathcal{D}} (a\bar{\sigma}_1(i) + b - \bar{\alpha}(i))^2. \end{aligned} \tag{52}$$

Now, we will prove that this result is not compatible with $\sigma_2 \sqsubset_{\mathbb{I}}^{\alpha} \sigma_1$, thus proving (47).

Among all the permissible transformation functions for the interval scale, let $f^*(\omega) = a^* \cdot \omega + b^*$ be the transformation that minimizes $\sum_{i \in \mathcal{D}} (f^*(\bar{\sigma}_2) - \bar{\alpha}(i))^2$. Let us denote with $\omega^* = f^*(\omega)$. Then, according to (52):

$$\begin{aligned} \sum_{i \in \mathcal{D}} (\bar{\sigma}_2^*(i) - \bar{\alpha}(i))^2 &= \sum_{i \in \mathcal{D}} (a^* \bar{\sigma}_2(i) + b^* - \bar{\alpha}(i))^2 \\ &\geq \sum_{i \in \mathcal{D}} (a^* \bar{\sigma}_1(i) + b^* - \bar{\alpha}(i))^2 = \sum_{i \in \mathcal{D}} (\bar{\sigma}_1^*(i) - \bar{\alpha}(i))^2. \end{aligned}$$

Therefore, the Euclidean distance to $\bar{\alpha}$ of the transformation of $\bar{\sigma}_1$ (i.e., $f^*(\bar{\sigma}_1)$) is smaller than the minimal Euclidean distance to $\bar{\alpha}$ of $\bar{\sigma}_2$ transformations. Now, given that two equivalent assignments belong to the same equivalence class ($\omega \approx_{\mathbb{T}} \omega' \implies [\omega] = [\omega']$) and they have the same set of permissible transformations, and given that $\bar{\sigma}_1 \approx_{\mathbb{I}} \sigma_1$ and $\bar{\sigma}_2 \approx_{\mathbb{I}} \sigma_2$, then we can say that the Euclidean distance to $\bar{\alpha}$ of the transformation of σ_1 (i.e., $f^*(\sigma_1)$) is lower than the minimal Euclidean distance to $\bar{\alpha}$ of σ_2 transformations:

$$\sum_{i \in \mathcal{D}} (\sigma_2^*(i) - \bar{\alpha}(i))^2 \geq \sum_{i \in \mathcal{D}} (\sigma_1^*(i) - \bar{\alpha}(i))^2. \tag{53}$$

Now, we have to prove that if an assignment ω_2 is value-closer to α than another assignment ω_1 , then the Euclidean distance must be lower:

$$\omega_2 \triangleleft_{\mathbb{I}}^{\alpha} \omega_1 \implies \sum (\omega_2 - \alpha)^2 < \sum (\omega_1 - \alpha)^2. \tag{54}$$

According to (6):

$$\omega_1 \triangleleft_I^\alpha \omega_2 \iff \forall d \in \mathcal{D}(\omega_1(d) \triangleleft_I^{\alpha(d)} \omega_2(d)) \wedge \exists d \in \mathcal{D}(\omega_1(d) \triangleleft_I^{\alpha(d)} \omega_2(d)).$$

According to Lemma 1, the consequent is equivalent to:

$$\begin{aligned} &\forall d \in \mathcal{D}(|\omega_1(d) - \alpha(d)| \leq |\omega_2(d) - \alpha(d)|) \\ &\wedge \exists d \in \mathcal{D}(|\omega_1(d) - \alpha(d)| < |\omega_2(d) - \alpha(d)|) \\ \implies &\forall d \in \mathcal{D}((\omega_1(d) - \alpha(d))^2 \leq (\omega_2(d) - \alpha(d))^2) \\ &\wedge \exists d \in \mathcal{D}((\omega_1(d) - \alpha(d))^2 < (\omega_2(d) - \alpha(d))^2) \\ \implies &\sum (\omega_2 - \alpha)^2 < \sum (\omega_1 - \alpha)^2. \end{aligned}$$

In addition, (54) can be also expressed as:

$$\sum (\omega_2 - \alpha)^2 \geq \sum (\omega_1 - \alpha)^2 \implies \neg(\omega_2 \triangleleft_I^\alpha \omega_1).$$

Therefore, considering (53) we have:

$$\neg(\sigma_2^* \triangleleft_I^{\bar{\alpha}} \sigma_1^*).$$

In other words, there does not exist any transformation of σ_2 strictly value-closer to $\bar{\alpha}$ than $f^*(\bar{\sigma}_1)$. Therefore, σ_2 is not equivalence-closer to $\bar{\alpha}$ than σ_1 , that is $\neg(\sigma_2 \sqsubset_I^{\bar{\alpha}} \sigma_1)$.

As the last step of the proof, we have to consider that

$$\sigma_2 \sqsubset_I^{\bar{\alpha}} \sigma_1 \iff \sigma_2 \sqsubset_I^\alpha \sigma_1.$$

which was already proved in the proof of Theorem 3, see (44) (indeed that proof is for non-strict closeness, but nothing changes for the strict version).

Therefore, we have proved (47) and thus (46). □

Proof of Theorem 13 We want to prove that the cosine is an equivalence-oriented metric for the ratio scale type, i.e., it satisfies EOI_R and EOM_R . The proof has the same structure of the previous proof of Theorem 12, but we repeat everything to make this proof self contained. A first remark is that the invariance under interval permissible transformation functions (i.e., ratio functions $y = a \cdot x$) is a well known property of cosine. Therefore, EOI_R is satisfied.

Then, we need to prove that cosine satisfies EOM at ratio scale type. That is, cosine with a gold α increases when going from an assignment σ_1 to an equivalence-closer to α assignment σ_2 :

$$\sigma_2 \sqsubset_R^\alpha \sigma_1 \implies \text{COS}(\sigma_2, \alpha) > \text{COS}(\sigma_1, \alpha). \tag{55}$$

We prove it by contraposition; therefore, we need to prove:

$$\text{COS}(\sigma_2, \alpha) \leq \text{COS}(\sigma_1, \alpha) \implies \neg(\sigma_2 \sqsubset_R^\alpha \sigma_1), \tag{56}$$

i.e., that if $\text{COS}(\sigma_2, \alpha) \leq \text{COS}(\sigma_1, \alpha)$ then σ_2 is not equivalence-closer to α than σ_1 .

We start by considering the normalization $\bar{\sigma}$ of an assignment σ as the ratio transformation:

$$\bar{\sigma}(i) = \frac{\sigma(i)}{\sqrt{\sum_{i \in \mathcal{D}} \sigma(i)^2}}$$

This normalization has the following properties:

$$\sum_{i \in \mathcal{D}} \bar{\sigma}(i) = \frac{\sum_{i \in \mathcal{D}} \sigma(i)}{\sqrt{\sum_{i \in \mathcal{D}} \sigma(i)^2}}, \tag{57}$$

$$\sum_{i \in \mathcal{D}} \bar{\sigma}(i)^2 = 1, \tag{58}$$

and

$$\text{COS}(\bar{\sigma}, \bar{\alpha}) = \frac{\sum_{i \in \mathcal{D}} \bar{\sigma}(i)\bar{\alpha}(i)}{\sqrt{\sum_{i \in \mathcal{D}} \bar{\sigma}(i)^2} \cdot \sqrt{\sum_{i \in \mathcal{D}} \bar{\alpha}(i)^2}} = \sum_{i \in \mathcal{D}} \bar{\sigma}(i)\bar{\alpha}(i) \tag{59}$$

Given that cosine is invariant under ratio permissible transformations, it holds that:

$$\text{COS}(\sigma_2, \alpha) \leq \text{COS}(\sigma_1, \alpha) \iff \text{COS}(\bar{\sigma}_2, \bar{\alpha}) \leq \text{COS}(\bar{\sigma}_1, \bar{\alpha}),$$

Then, according to (59)

$$\begin{aligned} \text{COS}(\bar{\sigma}_2, \bar{\alpha}) \leq \text{COS}(\bar{\sigma}_1, \bar{\alpha}) &\iff \sum_{i \in \mathcal{D}} \bar{\sigma}_2(i)\bar{\alpha}(i) \leq \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i)\bar{\alpha}(i) \\ &\iff -2 \sum_{i \in \mathcal{D}} \bar{\sigma}_2(i)\bar{\alpha}(i) \geq -2 \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i)\bar{\alpha}(i). \end{aligned} \tag{60}$$

Now, let us consider any linear transformation σ'_2 of the normalized assignment $\bar{\sigma}_2$. That is: $\sigma'_2(i) = a \cdot \bar{\sigma}_2(i)$ with $a > 0$. Then:

$$\begin{aligned} \sum_{i \in \mathcal{D}} (\sigma'_2(i) - \bar{\alpha}(i))^2 &= \sum_{i \in \mathcal{D}} (a\bar{\sigma}_2(i) - \bar{\alpha}(i))^2 \\ &= a^2 \sum_{i \in \mathcal{D}} \bar{\sigma}_2(i)^2 + \sum_{i \in \mathcal{D}} \bar{\alpha}(i)^2 - 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_2(i)\bar{\alpha}(i). \end{aligned}$$

According to (58):

$$= a^2 + \sum_{i \in \mathcal{D}} \bar{\alpha}(i)^2 - 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_2(i)\bar{\alpha}(i).$$

Then, according to (60) and given that $a > 0$:

$$\begin{aligned} &\geq a^2 + \sum_{i \in \mathcal{D}} \bar{\alpha}(i)^2 - 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i)\bar{\alpha}(i) \\ &= a^2 \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i)^2 + \sum_{i \in \mathcal{D}} \bar{\alpha}(i)^2 - 2a \sum_{i \in \mathcal{D}} \bar{\sigma}_1(i)\bar{\alpha}(i) \\ &= \sum_{i \in \mathcal{D}} (a\bar{\sigma}_1(i) - \bar{\alpha}(i))^2. \end{aligned}$$

That is, reading the equality/inequality chain, we can assert that for every value $a > 0$:

$$\text{COS}(\sigma_2, \alpha) \leq \text{COS}(\sigma_1, \alpha) \implies \sum_{i \in \mathcal{D}} (a\bar{\sigma}_2(i) - \bar{\alpha}(i))^2 \geq \sum_{i \in \mathcal{D}} (a\bar{\sigma}_1(i) - \bar{\alpha}(i))^2. \tag{61}$$

Now, we will prove that this result is not compatible with $\sigma_2 \sqsubset_{\mathbb{R}}^{\alpha} \sigma_1$, thus proving (56).

Among all the permissible transformation functions for the interval scale, let $f^*(\omega) = a^* \cdot \omega$ be the transformation that minimizes $\sum_{i \in \mathcal{D}} (f^*(\bar{\sigma}_2) - \bar{\alpha}(i))^2$. Let us denote with $\omega^* = f^*(\omega)$. Then, according to (61):

$$\sum_{i \in \mathcal{D}} (\bar{\sigma}_2^*(i) - \bar{\alpha}(i))^2 = \sum_{i \in \mathcal{D}} (a^* \bar{\sigma}_2(i) - \bar{\alpha}(i))^2 \geq \sum_{i \in \mathcal{D}} (a^* \bar{\sigma}_1(i) - \bar{\alpha}(i))^2 = \sum_{i \in \mathcal{D}} (\bar{\sigma}_1^*(i) - \bar{\alpha}(i))^2.$$

Therefore, the Euclidean distance to $\bar{\alpha}$ of the transformation of $\bar{\sigma}_1$ (i.e., $f^*(\bar{\sigma}_1)$) is smaller than the minimal Euclidean distance to $\bar{\alpha}$ of $\bar{\sigma}_2$ transformations. Now, given that two equivalent assignments belong to the same equivalence class ($\omega \approx_{\mathbb{T}} \omega' \implies [\omega] = [\omega']$) and they have the same set of permissible transformations, and given that $\bar{\sigma}_1 \approx_{\mathbb{T}} \sigma_1$ and $\bar{\sigma}_2 \approx_{\mathbb{T}} \sigma_2$, then we can say that the Euclidean distance to $\bar{\alpha}$ of the transformation of σ_1 (i.e., $f^*(\sigma_1)$) is lower than the minimal Euclidean distance to $\bar{\alpha}$ of σ_2 transformations:

$$\sum_{i \in \mathcal{D}} (\sigma_2^*(i) - \bar{\alpha}(i))^2 \geq \sum_{i \in \mathcal{D}} (\sigma_1^*(i) - \bar{\alpha}(i))^2. \tag{62}$$

Now, we have to prove that if an assignment ω_2 is value-closer to α than another assignment ω_1 , then the Euclidean distance must be lower:

$$\omega_2 \triangleleft_{\mathbb{R}}^{\alpha} \omega_1 \implies \sum (\omega_2 - \alpha)^2 < \sum (\omega_1 - \alpha)^2. \tag{63}$$

According to (6):

$$\omega_1 \triangleleft_{\mathbb{R}}^{\alpha} \omega_2 \iff \forall d \in \mathcal{D} \left(\omega_1(d) \leq_{\mathbb{R}}^{\alpha(d)} \omega_2(d) \right) \wedge \exists d \in \mathcal{D} \left(\omega_1(d) <_{\mathbb{R}}^{\alpha(d)} \omega_2(d) \right).$$

According to Lemma 1 the consequent is equivalent to:

$$\begin{aligned} & \forall d \in \mathcal{D} (|\omega_1(d) - \alpha(d)| \leq |\omega_2(d) - \alpha(d)|) \\ & \wedge \exists d \in \mathcal{D} (|\omega_1(d) - \alpha(d)| < |\omega_2(d) - \alpha(d)|) \\ \implies & \forall d \in \mathcal{D} ((\omega_1(d) - \alpha(d))^2 \leq (\omega_2(d) - \alpha(d))^2) \\ & \wedge \exists d \in \mathcal{D} ((\omega_1(d) - \alpha(d))^2 < (\omega_2(d) - \alpha(d))^2) \\ \implies & \sum (\omega_2 - \alpha)^2 < \sum (\omega_1 - \alpha)^2. \end{aligned}$$

In addition, (63) can be also expressed as:

$$\sum (\omega_2 - \alpha)^2 \geq \sum (\omega_1 - \alpha)^2 \implies \neg(\omega_2 \triangleleft_{\mathbb{R}}^{\alpha} \omega_1).$$

Therefore, considering (62) we have:

$$\neg(\sigma_2^* \triangleleft_{\mathbb{R}}^{\bar{\alpha}} \sigma_1^*).$$

In other words, there does not exist any transformation of σ_2 strictly value-closer to $\bar{\alpha}$ than $f^*(\sigma_1)$. Therefore, σ_2 is not equivalence-closer to $\bar{\alpha}$ than σ_1 , that is $\neg(\sigma_2 \sqsubset_{\mathbb{R}}^{\bar{\alpha}} \sigma_1)$.

As the last step of the proof, we have to consider that

$$\sigma_2 \sqsubset_{\mathbb{R}}^{\bar{\alpha}} \sigma_1 \iff \sigma_2 \sqsubset_{\mathbb{R}}^{\alpha} \sigma_1,$$

which was already proved in the proof of Theorem 3, see (44) (indeed that proof is for non-strict closeness, but nothing changes for the strict version). Therefore, we have proved (56) and thus (55). \square

References

- Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., et al. (2015). SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on interpretability. In *Proceedings of SemEval, 2015* (pp. 252–263).
- Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., et al. (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of SemEval, 2016* (pp. 509–523).
- Amigo, E., Gonzalo, J., Verdejo, F., & Spina, D. (2019). A comparison of filtering evaluation metrics based on formal constraints. *Information Retrieval Journal*, 22(6), 581–619.
- Amigó, E., Gonzalo, J., Artiles, J., & Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4), 461–486.
- Amigó, E., Gonzalo, J., & Mizzaro, S. (2014). A general account of effectiveness metrics for information tasks: retrieval, filtering, and clustering. In *Proceedings of SIGIR* (pp. 1289–1289). ACM.
- Amigó, E., Gonzalo, J., & Mizzaro, S. (2015). A formal approach to effectiveness metrics for information access: Retrieval, filtering, and clustering. In: *ECIR 2015: Advances in information retrieval* (pp. 817–821).
- Amigó, E., Gonzalo, J., & Verdejo, F. (2011). A comparison of evaluation metrics for document filtering. In *CLEF, LNCS* (Vol. 6941, pp. 38–49). Springer.
- Amigó, E., Gonzalo, J., & Verdejo, F. (2013). A general evaluation measure for document organization tasks. In *Proceedings of SIGIR* (pp. 643–652). ACM, New York, NY, USA.
- Amigó, E., Gonzalo, J., Spina, D., & Verdejo, F. (2018). A comparison of filtering evaluation metrics based on formal constraints. *Information Retrieval*. (in press).
- Baccianella, S., Esuli, A., & Sebastiani, F. (2009). Evaluation measures for ordinal regression. In *Proceedings of the 2009 ninth international conference on intelligent systems design and applications, ISDA '09* (pp. 283–287).
- Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., & Patti, V. (2016). Overview of the Evalita 2016 SENTiment POLarity classification task. In *Proceedings of third Italian conference on computational linguistics (CLiC-it 2016)*.
- Bollmann, P. (1984). Two axioms for evaluation measures in information retrieval. In *SIGIR '84* (pp. 233–245). Swinton: British Computer Society.
- Busin, L., & Mizzaro, S. (2013). Axiometrics: An axiomatic approach to information retrieval effectiveness metrics. In *Proceedings of ICTIR 2013* (pp. 22–29). ACM.
- Cardoso, J. S., & Sousa, R. (2011). Measuring the performance of ordinal classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(08), 1173–1195. <https://doi.org/10.1142/S0218001411009093>.
- Della Mea, V., & Mizzaro, S. (2004). Measuring retrieval effectiveness: A new proposal and a first experimental validation. *Journal of the American Society for Information Science*, 55(6), 530–543.
- Dom, B. (2001). An information-theoretic external cluster-validity measure. IBM Research Report. <http://citeseer.ist.psu.edu/dom01informationtheoretic.html>.
- Ferrante, M., Ferro, N., & Maistro, M. (2015). Towards a formal framework for utility-oriented measurements of retrieval effectiveness. In *Proceedings of ICTIR* (pp. 21–30).
- Ferrante, M., Ferro, N., & Pontarollo, S. (2017). Are IR evaluation measures on an interval scale? In *Proceedings of ACM ICTIR*. (pp. 67–74). ACM.
- Ferrante, M., Ferro, N., & Pontarollo, S. (2019). A general theory of IR evaluation measures. *IEEE Transactions on Knowledge & Data Engineering*, 31(3), 409–422.
- Ferri, C., Hernández-Orallo, J., & Modroui, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1), 27–38.

- Fuhr, N. (2018). Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51(3), 32–41.
- Gaudette, L., & Japkowicz, N. (2009). Evaluation methods for ordinal classification. In Y. Gao & N. Japkowicz (Eds.), *Advances in Artificial Intelligence. Canadian AI 2009. Lecture Notes in Computer Science* (Vol. 5549). Berlin, Heidelberg: Springer.
- González, P., Castaño, A., Chawla, N. V., & Coz, J. J. D. (2017). A review on quantification learning. *ACM Computing Surveys*, 50(5), 74:1–74:40.
- Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2–3), 107–145.
- Maddalena, E., & Mizzaro, S. (2014). Axiometrics: Axioms of information retrieval effectiveness metrics. In *Proceedings of the sixth EVIA workshop* (pp. 17–24).
- Maddalena, E., Mizzaro, S., Scholer, F., & Turpin, A. (2017). On crowdsourcing relevance magnitudes for information retrieval evaluation. *ACM Transactions on Information Systems (TOIS)*, 35(3), 19:1–19:32.
- Meila, M. (2003). Comparing clusterings. In *Proceedings of COLT 03*.
- Moffat, A. (2013). Seven numeric properties of effectiveness metrics. In *AIRS'13* (pp. 1–12).
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: an integrated approach*. Newark: Lawrence Erlbaum Associates.
- Qi, H., Yang, M., He, X., & Li, S. (2010). Re-examination on Lam% in spam filtering. In *Proceedings of SIGIR*.
- Reimers, N., Beyer, P., & Gurevych, I. (2016). Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING, 2016* (pp. 87–96).
- Roberts, F. (1984). *Measurement theory: volume 7: with applications to decisionmaking, utility, and the social sciences. Encyclopedia of mathematics and its applications*. Cambridge: Cambridge University Press.
- Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of EMNLP-CoNLL* (pp. 410–420).
- Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of SemEval '17*. ACL.
- Sebastiani, F. (2015). An axiomatically derived measure for the evaluation of classification algorithms. In *Proceedings of ICTIR 2015* (pp 11–20). ACM.
- Sokolova, M. (2006). Assessing invariance properties of evaluation measures. In *Proceedings of NIPS'06 workshop on testing deployable learning and decision systems*
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–80.
- Suppes, P., & Zinnes, J. L. (1963). *Basic measurement theory. Handbook of mathematical psychology* (Vol. 1, pp. 3–76). New York: Wiley.
- Van Rijsbergen, K. (1974). Foundation of evaluation. *Journal of Documentation*, 30(4), 365–373.
- van Rijsbergen, K. (1981). *Retrieval effectiveness* (pp. 32–43). London: Butterworths. (chap 3).
- Wu, W., Xiong, H., & Shekhar, S. (Eds.). (2003). *Clustering and information retrieval*. Alphen aan den Rijn: Kluwer.
- Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of SIGIR* (pp. 267–273). ACM.
- Yao, Y. (1995). Measuring retrieval effectiveness based on user preference of documents. *JASIS*, 46, 133–145.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.