Statistical Performance Metrics for Use with Imprecise Ground-Truth

Bart Lamiroy, Pascal Pierrot

Statistical Performance Metrics for Use with Imprecise Ground-Truth

Bart Lamiroy[1] and Pascal Pierrot[2]

[1] Université de Lorraine – Loria (UMR 7503)
Campus Scientifique – BP 239
54506 Vandœuvre-lès-Nancy cedex – France
bart.lamiroy@loria.fr
[2] Université de Lorraine – Mines Nancy
Campus Artem - CS 14 234
92 Rue Sergent Blandan
54042 Nancy – France

Abstract. This paper addresses performance evaluation in the presence of imprecise ground-truth. Indeed, the most common assumption when performing benchmarking measures is that the reference data is flawless. In previous work, we have shown that this assumption cannot be taken for granted, and that, in the case of perceptual interpretation problems it is most certainly always wrong but for the most trivial cases.

We are presenting a statistical test that will allow measuring the confidence one can have in the results of a benchmarking test ranking multiple algorithms. More specifically, we can express the probability of the ranking not being respected in the presence of a given level of errors in the ground truth data.

1 Introduction

In this paper we investigate statistical tests for assessing the risk of misranking algorithms on benchmarks when using unreliable ground truth. The current approach to performance analysis is that algorithms assumed to be tested on totally reliable ground truth. We have shown in previous work that this assumption is flawed, and that there is an inherent interpretative bias in the definition of ground truth.

This paper is a first tentative step in creating a mathematically sound framework for assessing the risk of relying on imprecise ground truth. Indeed, the probability of algorithms being misranked is directly dependent on their overall performance on the one hand, and the level of error in the used ground truth.

This framework can be applied to benchmarking and contests (*e.g.* the GREC Arc Detection contests [7, 1–3] or more general benchmarking environments [6])

2 Problem Description

2.1 Definitions and Notations

In this section we introduce all definitions and notations we use throughout this paper.

Let $\Phi = \{\phi_1, ..., \phi_p\}$ be a set of data, $\mathcal{I} = \{i_1, ..., i_q\}$ a finite set of possible interpretations over Φ and $\mathsf{A} = \{A_1, ..., A_n\}$ a set of algorithms.

First, we define the notion of *Ground Truth*, which associates a truth value to the interpretation i of a given data element d.

Definition 1 (Ground Truth) *A ground truth is a function Ω such that:*

$$\Omega : \Phi \times \mathcal{I} \to \{0, 1\}$$
$$(\phi, i) \mapsto \begin{cases} 1 \text{ iff } i \textit{ is a correct interpretation for } \phi \\ 0 \textit{ otherwise} \end{cases}$$

Definition 2 (Algorithm) *An algorithm A is a function associating one or multiple interpretations to a given data element ϕ.*

$$A : \Phi \to \{0, 1\}^q$$
$$\phi \mapsto (a_1, ..., a_q)$$
$$\textit{with } a_k\textit{=1 if } \phi \textit{ has } i_k \in \mathcal{I} \textit{ as interpretation}$$
$$\textit{and } a_k\textit{=0 otherwise}$$

We use the following notation, for $k \in \{1, ..., p\}$: $A_j(\phi_k) = (a^j_{k1}, ..., a^j_{kq})$, as shown in Table 1

	A_1				...	A_n			
	i_1	i_2	...	i_q	...	i_1	i_2	...	i_q
ϕ_1	a^1_{11}	a^1_{12}		a^1_{1q}		a^n_{11}	a^n_{12}		a^n_{1q}
ϕ_2		a^1_{23}	...				a^n_{23}	...	
...									
ϕ_p	a^1_{p1}								a^n_{pq}

Table 1. Example representation of data, algorithms and interpretations

2.2 Algorithm Ranking

Performance analysis is generally done by ranking algorithms with respect to their results on ground truth data. In order to correctly establish the claims made in this paper, we need to formalise the notion of ranking order of algorithms with respect to a given ground truth.

	A_1				A_2				A_3				Ω			
	i_1	i_2	i_3	i_4	i_1	i_2	i_3	i_4	i_1	i_2	i_3	i_4	i_1	i_2	i_3	i_4
ϕ_1	0	1	0	1	0	1	**1**	1	0	1	0	1	0	1	0	1
ϕ_2	0	0	**1**	**0**	0	0	0	1	**1**	0	0	1	0	0	0	1
ϕ_3	0	1	1	0	0	1	1	0	0	1	**0**	0	0	1	1	0
ϕ_4	0	1	**1**	0	0	1	0	0	1	**1**	0	0	0	1	0	0
ϕ_5	1	0	0	0	1	0	0	0	1	**1**	0	0	1	0	0	0

Table 2. Example of algorithms performing against a perfect ground truth Ω

Definition 3 (Ranking preorder) *A ranking preorder is expressed with respect to a ground truth Ω, and is defined for set of algorithms* $\mathbf{A}$*, a data set Φ and a set of interpretations $\mathcal{I}$.*

We note $\prec_\Omega$ a preorder on $\mathbf{A}$ *such that $A_1 \prec_\Omega A_2$* iff

$$\left|\left\{(k,l)|a^1_{k,l} = \Omega(\phi_k, i_l)\right\}\right| \leq \left|\left\{(k,l)|a^2_{k,l} = \Omega(\phi_k, i_l)\right\}\right|$$

In other terms, algorithms are compared with respect to the cardinality of their agreement with the ground truth.

3 Performance Metrics on Flawless Ground Truth

Comparing algorithms in the presence of perfectly reliable ground truth is what is usually practiced, and does not require extensive statistical approaches, as we recall here. Assuming Ω represents a 100% reliable ground truth, Table 2 shows an example of algorithm outputs for a set of data items.

Following the formalism in Definition 3 the most straightforward approach for ranking is to compute the percentage of good answers of each algorithm and to rank them accordingly.

The proportion of correct answers for algorithm A_j is:

$$\tau_{A_j} = \frac{\sum_{\substack{k\in\{1,\ldots,p\}\\ l\in\{1,\ldots,q\}}} \delta_{a^j_{k,l},\Omega(\phi_k,i_l)}}{pq} \tag{1}$$

$$\text{where} \quad \delta_{i,j} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases} \tag{2}$$

Using τ_{A_j} we can instantiate the algorithm ranking above such that $A_i \prec_\Omega A_j$ *iff* $\tau_{A_i} <= \tau_{A_j}$ (the better ranked algorithms are on the right, the less ranked on the left).

Applied to the data in Table 2, we obtain $\tau_{A_1} = 0.85$, $\tau_{A_2} = 0.95$, $\tau_{A_3} = 0.8$ and therefore $A_3 \prec_\Omega A_1 \prec_\Omega A_2$.

4 Performance Metrics on Flawed Ground Truth

As we have shown in [5] it is virtually impossible to obtain "error-free" ground truth. We therefore assume Ω contains a proportion of incorrect data. We also assume this proportion is bound by a known value $\varepsilon \in [0, 1]$.

4.1 General Approach

We would like to know how reliable the ranking of a set of algorithms is when it is based on a ground truth that is reliable up to ε. In order to achieve this, let $\overline{\Omega}$ be the absolute ground truth, to which we have no access other than its approximation Ω_ε.

In other terms,

$$\frac{\left|\left\{(\phi, i) \in \Phi \times \mathcal{I} \,\middle|\, \Omega_\varepsilon(\phi, i) \neq \overline{\Omega}(\phi, i)\right\}\right|}{|\Phi|\,|\mathcal{I}|} \leq \varepsilon$$

Given Ω_ε we can use Definition 3 to define $\prec_{\Omega_\varepsilon}$. The question is whether it is possible to determine if this ranking is a reliable approximation of $\prec_{\overline{\Omega}}$ to which we have no access.

The rest of this paper will address the various probabilistic approaches that will allow us to quantify this difference.

4.2 Notations

In order to develop probabilistic methods, we define the following random variables:

For a given ground truth Ω, an algorithm A_j with $j \in \{1, ..., n\}$, $k \in \{1, ..., p\}$ and $l \in \{1, ..., q\}$, the random variable $X_{k,l}^{\Omega,j}$ expresses whether algorithm A_j correctly interprets ϕ_k as i_l.

This entails that

$$X_{k,l}^{\Omega,j} = \begin{cases} 1 \text{ if } \; A_j(\phi_k)|_l = \Omega(\phi_k, i_l) \\ 0 \text{ otherwise} \end{cases}$$

which can be expressed also as

$$X_{k,l}^{\Omega,j} = \begin{cases} 1 \text{ if } a_{kl}^j = \Omega(\phi_k, i_l) \\ 0 \text{ otherwise} \end{cases}$$

the realisations of which will be noted as $x_{k,l}^{\Omega,j}$.

The associated probabilities will be noted as

$$\mathbf{P}\left(X_{k,l}^{\Omega,j} = 1\right) = p_{k,l}^{\Omega,j}$$

and

$$\mathbf{P}\left(X_{k,l}^{\Omega,j} = 0\right) = 1 - p_{k,l}^{\Omega,j}$$

with $p_{k,l}^j \in [0, 1]$ $\forall i, j, k$.

5 Simplified Approach Using Two Algorithms

Given the impossibility of using χ^2 multinomial adequacy testing (since it requires knowledge on $\overline{\Omega}$, which we don't have), we have tried to formally and numerically derive statistics on simplified data. In this section we restrict ourselves to two algorithms and one single possible interpretation per algorithm. This will allow us to establish a first category of statistical tests.

	A_1	A_2	Ω_ε	$\overline{\Omega}$
	i	i	i	i
ϕ_1	0	1	0	$x_1^{\overline{\Omega}}$
ϕ_2	0	1	0	$x_2^{\overline{\Omega}}$
ϕ_3	0	1	0	$x_3^{\overline{\Omega}}$
ϕ_4	0	1	0	$x_4^{\overline{\Omega}}$

5.1 Working Hypotheses and Notations

We now assume there are only two algorithms A_1 and A_2 to be compared. We are assuming these algorithms are binary classifiers (*i.e.* there is only one possible interpretation: $\mathcal{I} = \{i\}$; algorithms categorize data in `true` or `false`)

Consequently, we denote the set of values of an algorithm over Φ as $\{a_k^j\}$ and the set of corresponding ground truth values as $\Omega_{cvarepsilon}(\phi_k)$. Once again, let $\overline{\Omega}$ be a perfect ground truth, for which we have no a priori knowledge, nor any algorithm ranking. Similarly, Ω_ε is a ground truth for which all values are known, and of which we know that it differs from $\overline{\Omega}$ by at most ε. This allows us to express a ranking order $\prec_{\Omega_\varepsilon}$ between A_1 and A_2.

In order to develop the rest of our rationale, we need to introduce the notion of *divergence* between two algorithms A_1 and A_2:

Definition 4 (Disagreement Set) *Let A_1 and A_2 be two algorithms of which the results to a known ground truth Ω are known. These results are respectively $a_{k,l}^{\Omega,1}$ and $a_{k,l}^{\Omega,2}$ with $k \in \{1,...,p\}, l \in \{1,...,q\}$.*

We define the disagreement set between A_1 and A_2 as

$$\mathcal{D}(A_1, A_2) = \left\{ (k,l) \,\middle|\, a_{k,l}^{\Omega,1} \neq a_{k,l}^{\Omega,2} \right\}.$$

We can extend this notation to also express the disagreement between an algorithm and the ground truth, or between ground truths:

$$\mathcal{D}\left(A_i, \Omega\right) = \left\{ (k,l) \,\middle|\, a_{k,l}^{\Omega,i} \neq x_{k,l}^{\Omega} \right\}$$
$$\mathcal{D}\left(\overline{\Omega}, \Omega_\varepsilon\right) = \left\{ (k,l) \,\middle|\, x_{k,l}^{\overline{\Omega}} \neq x_{k,l}^{\Omega_\varepsilon} \right\}$$

Definition 5 (Agreement Set) *Let A_1 and A_2 be two algorithms of which the results to a known ground truth Ω are known. These results are respectively $a_{k,l}^{\Omega,1}$ and $a_{k,l}^{\Omega,2}$ with $k \in \{1,...,p\}, l \in \{1,...,q\}$.*

We define the agreement set between A_1 and A_2 as

$$\mathcal{A}(A_1, A_2) = \left\{ (k,l) \,\middle|\, a_{k,l}^{\Omega,1} = a_{k,l}^{\Omega,2} \right\}.$$

It is straightforward to note that $\mathcal{A}$ and $\mathcal{D}$ are complements of each other:

$$\mathcal{A}(X,Y) = \overline{\mathcal{D}(X,Y)}.$$

Definition 6 (Divergence between two algorithms) *Let A_1 and A_2 be two algorithms. Given their disagreement sets as per Definition 4, we define divergence between A_1 and A_2 as*

$$\mathbf{D}(A_1, A_2) = |\mathcal{D}(A_1, A_2)|.$$

Divergence is equivalent to the Hamming distance between the vectors containing the output values of A_1 and A_2.

As in the case of the disagreement set, this definition can be extended to express the difference of agreement between algorithms and the ground truth, or between ground truths:

$$\mathbf{D}(A_i, \Omega) = |\mathcal{D}(A_i, \Omega)|$$
$$\mathbf{D}\left(\overline{\Omega}, \Omega_\varepsilon\right) = \left|\mathcal{D}(\overline{\Omega}, \Omega)\right|$$

5.2 Divergence Estimation

It is straightforward to prove that, in the general case, given any ground truth Ω, the following in equations hold:

$$\mathbf{D}(A_1, A_2) \leq \mathbf{D}(A_1, \Omega) + \mathbf{D}(A_2, \Omega) \tag{3}$$
$$\mathbf{D}(A_2, \Omega) - \mathbf{D}(A_1, \Omega) \leq \mathbf{D}(A_1, A_2) \tag{4}$$

Explanation (3) results from the triangular inequality of the Hamming distance. It can also be explained by the fact that, on binary classifiers, at best, two algorithms disagree with the ground truth on data points on which they disagree with one another. At worst, both algorithms perfectly agree on all data points (even when in disagreement with the ground truth), in which case $\mathbf{D}(A_1, A_2) = 0$.

(4) can be derived from the fact that the difference in disagreement of two algorithms with the ground truth is at worst their disagreement with one another. This can be easily observed from the extreme configurations where either $\mathbf{D}(A_1, A_2) = 0$ (and consequently $\mathbf{D}(A_2, \Omega) = \mathbf{D}(A_1, \Omega)$) or either $\mathbf{D}(A_1, A_2) = \mathbf{D}(A_1, \Omega) + \mathbf{D}(A_2, \Omega)$.

Given the fact that the divergence between $\overline{\Omega}$ and Ω_ε is bounded by $\varepsilon\, p$ we can deduce that

$$\mathbf{D}(A_i, \overline{\Omega}) - \varepsilon\, p \leq \mathbf{D}(A_i, \Omega_\varepsilon) \leq \mathbf{D}(A_i, \overline{\Omega}) + \varepsilon\, p \tag{5}$$

This implies, by combining (3) and (5), that

$$\mathbf{D}(A_1, \overline{\Omega}) \leq \mathbf{D}(A_1, A_2) + \mathbf{D}(A_2, \Omega_\varepsilon) + \varepsilon\, p \tag{6}$$

Similarly, by combining (4) and (5), we get

$$\mathbf{D}(A_1, \overline{\Omega}) \geq \mathbf{D}(A_1, A_2) - \mathbf{D}(A_2, \Omega_\varepsilon) - \varepsilon\, p \tag{7}$$

Therefore, and because of the symmetry of the demonstration, the divergence between any given algorithm with the (unknown) perfect ground truth, is bounded by the (known) disagreement between both algorithms and the assumed error level of the (known) tainted ground truth.

5.3 Estimating the Probability of a Change in Ranking

At this point in the process, we have a tainted ground truth Ω_ε and two algorithms A_1 and A_2, which we can rank using $\prec_{\Omega_\varepsilon}$. The question that arises is that, if we had done the ranking based on $\overline{\Omega}$ (*i.e.* without the ε ground truth error), would the ordering of the two algorithms have changed?

We therefore address the question of whether a change in the value of ε changes the ranking of the algorithms, by formalising it as a probabilistic problem. Let $\mathbf{P}\left(A_1 \prec_{\overline{\Omega}} A_2 \,\middle|_{A_1 \prec_{\Omega_\varepsilon} A_2}\right)$ be the probability that the ranking remains unchanged for ground truth $\overline{\Omega}$ given the ranking for Ω_ε.

Let us consider the specific example where $|\mathbf{D}(A_1, \Omega_\varepsilon) - \mathbf{D}(A_2, \Omega_\varepsilon)| > 2\varepsilon p$ (p being the number of data elements in Φ). In this case, the order $\prec_{\Omega_\varepsilon}$ on A_1 and A_2 will be strictly equivalent to $\prec_{\overline{\Omega}}$.

Indeed, let's assume $A_1 \prec_{\Omega_\varepsilon} A_2$. In the worst case, both A_1 and A_2 agree with Ω_ε on those data where Ω_ε gets it wrong with respect to $\overline{\Omega}$. In that situation, the following holds:

$$\mathcal{D}\left(\overline{\Omega}, \Omega_\varepsilon\right) \subset \mathcal{D}\left(A_1, \overline{\Omega}\right) \cap \mathcal{D}\left(A_2, \overline{\Omega}\right)$$

where, by definition, for the worst case scenario, $\mathbf{D}\left(\overline{\Omega}, \Omega_\varepsilon\right) = p\varepsilon$. Therefore

$$\mathbf{D}(A_1, \overline{\Omega}) = \mathbf{D}(A_1, \Omega_\varepsilon) + p\varepsilon$$

and

$$\mathbf{D}(A_2, \overline{\Omega}) = \mathbf{D}(A_2, \Omega_\varepsilon) - p\varepsilon$$

Since we made the assumption that $A_1 \prec_{\Omega_\varepsilon} A_2$ (and thus $\mathbf{D}(A_1, \Omega_\varepsilon) - \mathbf{D}(A_2, \Omega_\varepsilon) > 0$), we obtain that $\mathbf{D}(A_1, \overline{\Omega}) - \mathbf{D}(A_2, \overline{\Omega}) > 0$, and consequently $A_1 \prec_{\overline{\Omega}} A_2$.

We can therefore conclude that

$$\mathbf{P}\left(A_1 \prec_{\overline{\Omega}} A_2 \,\middle|_{A_1 \prec_{\Omega_\varepsilon} A_2}\right) = 1$$

if $\mathbf{D}(A_2, \Omega_\varepsilon) - \mathbf{D}(A_1, \Omega_\varepsilon) \geq 2\varepsilon p$.

If we consider $\mathcal{A}\left(A_2, \Omega_\varepsilon\right)$ the set of values common to A_2 and Ω_ε, and $\mathcal{D}\left(A_1, \Omega_\varepsilon\right)$ the set values where A_1 and Ω_ε differ, then

$$\mathbf{P}\left(A_1 \prec_{\overline{\Omega}} A_2 \,\middle|_{A_1 \prec_{\Omega_\varepsilon} A_2}\right) = 1 \qquad \text{if}$$

$$\left|\mathcal{A}\left(A_2, \Omega_\varepsilon\right) \cap \mathcal{D}\left(A_1, \Omega_\varepsilon\right)\right| < \frac{\mathbf{D}(A_2, \Omega_\varepsilon) - \mathbf{D}(A_1, \Omega_\varepsilon)}{2}$$

Explanation The only case where A_1 could inverse its ranking with respect to A_2 is when there is a sufficiently high number of values in $\mathcal{A}(A_2, \Omega_\varepsilon) \cap \mathcal{D}(A_1, \Omega_\varepsilon)$ for which $\overline{\Omega}$ differs from Ω_ε. We need at least $\frac{\mathbf{D}(A_2,\Omega_\varepsilon)-\mathbf{D}(A_1,\Omega_\varepsilon)}{2}$.

In the general case, one can use a Monte-Carlo simulation [8] in which the values of $\Omega_\varepsilon(\phi_k)$ are changed with a probability of ε. Although there is a fundamental difference between having a strictly bounded ground truth uncertainty of ε and using the same value as the likelihood of individual values being different between $\overline{\Omega}$ and Ω_ε, we can use the approach for estimating $\mathbf{P}\left(A_1 \prec_{\overline{\Omega}} A_2 \,\middle|_{A_1 \prec_{\Omega_\varepsilon} A_2}\right)$. The main reasons are that the high number of iterations of the Monte-Carlo simulation will converge to a Gaussian distribution around ε errors, and that in real situations, ε is usually an approximate guess with an associated Gaussian uncertainty. Furthermore, the closed-form developments in the next section support the numerical simulations obtained here.

Algorithm Execute the following loop N times (for large values of N)

- all values k in $[1..p]$, change the value of $\Omega_\varepsilon(\phi_k)$ with a probability of ε;
- compute $\mathbf{D}(A_2, \Omega_\varepsilon) - \mathbf{D}(A_1, \Omega_\varepsilon)$;
- if $\mathbf{D}(A_2, \Omega_\varepsilon) - \mathbf{D}(A_1, \Omega_\varepsilon) \geq 0$, increase counter c by 1.

After N iterations $\frac{c}{N}$ yields an approximation of the required probability.

Example we have used a Matlab implementation of Monte-Carlo on the data in Table 3. In our example $A_1 \prec_{\Omega_\varepsilon} A_2$ and $p = 10$.

	A_1	A_2	Ω_ε	$\overline{\Omega}$
	i	i	i	i
ϕ_1	0	0	0	$x_1^{\overline{\Omega}}$
ϕ_2	1	1	1	$x_2^{\overline{\Omega}}$
ϕ_3	1	1	1	$x_3^{\overline{\Omega}}$
ϕ_4	1	0	0	$x_4^{\overline{\Omega}}$
ϕ_5	1	0	0	$x_5^{\overline{\Omega}}$
ϕ_6	0	1	1	$x_6^{\overline{\Omega}}$
ϕ_7	0	1	0	$x_7^{\overline{\Omega}}$
ϕ_8	0	0	1	$x_8^{\overline{\Omega}}$
ϕ_9	0	0	1	$x_9^{\overline{\Omega}}$
ϕ_{10}	1	1	0	$x_{10}^{\overline{\Omega}}$

Table 3. Simple example of two algorithms performing against binary ground truth.

With $N = 10^5$ tests and a confidence level of 95% we obtain:

- $\varepsilon = 0.5$: $\mathbf{P}(A_1 \prec_{\overline{\Omega}} A_2) = 0.3135 \pm 0.002876$

- $\varepsilon = 0.2$: $\mathbf{P}(A_1 \prec_{\overline{\Omega}} A_2) = 0.5893 \pm 0.003049$
- $\varepsilon = 0.1$: $\mathbf{P}(A_1 \prec_{\overline{\Omega}} A_2) = 0.7517 \pm 0.002677$
- $\varepsilon = 0$: as expected $\mathbf{P}(A_1 \prec_{\overline{\Omega}} A_2) = 1$

5.4 Formal Approach

It is to be noted that the previous Monte-Carlo based solution is a mere convenience, and that a formal solution can be derived as well.

Indeed, one can make the following observations:

1. Given the fact that Ω_ε differs from $\overline{\Omega}$ by ε, the probability of $\Omega_\varepsilon(\phi_i)$ differing from $\overline{\Omega}(\phi_i)$, expressed as $\mathbf{P}\left(x_i^{\overline{\Omega}} \neq x_i^{\Omega_\varepsilon}\right)$, can be considered to be following a Bernoulli law of parameter ε. (This is exactly what is expressed by the probability of change in the above Monte-Carlo approach)
2. We are trying to measure the impact of a disagreement between Ω_ε and $\overline{\Omega}$ on the ranking between A_1 and A_2. For each ϕ_i where A_1 and A_2 are in agreement (regardless whether they agree or not with Ω_ε) a change in $\Omega_\varepsilon(\phi_i)$ is not going to affect the ranking $A_1 \prec_{\Omega_\varepsilon} A_2$ since both will be affected in the same sense.
   As a consequence, $\mathbf{P}(A_1 \prec_{\overline{\Omega}} A_2)$ only depends on the probability of Ω_ε being in disagreement with $\overline{\Omega}$ on only those ϕ_i where A_1 disagrees with A_2.

Based on these observations, and using the same notations as before, we can compute the probability of $\mathbf{P}(A_1 \prec_{\overline{\Omega}} A_2)$ as follows.

Let $\mathcal{D}_{A_1}$ (resp. $\mathcal{D}_{A_2}$) be the subset of $\mathcal{D}(A_1, A_2)$ where A_1 (resp. A_2) is in agreement with Ω_ε while being in disagreement with A_2 (resp. A_1).

$$\mathcal{D}_{A_1} = \mathcal{D}(A_1, A_2) \cap \mathcal{A}(A_1, \Omega_\varepsilon)$$

$$\mathcal{D}_{A_2} = \mathcal{D}(A_1, A_2) \cap \mathcal{A}(A_2, \Omega_\varepsilon)$$

Since we are only considering two algorithms, and given observation 2 above, this is equivalent to

$$\mathcal{D}_{A_1} = \mathcal{D}(A_1, A_2) - \mathcal{D}(A_2, \Omega_\varepsilon)$$

$$\mathcal{D}_{A_2} = \mathcal{D}(A_1, A_2) - \mathcal{D}(A_1, \Omega_\varepsilon)$$

It is straightforward to prove that $\mathcal{D}_{A_1} \cap \mathcal{D}_{A_2} = \emptyset$ and that $A_1 \prec_{\Omega_\varepsilon} A_2$ *iff* $\mathbf{D}_{A_1} \leq \mathbf{D}_{A_2}$ (where $\mathbf{D}$ expresses the cardinality of $\mathcal{D}$).

Furthermore, $\mathcal{D}_{A_1} \cap \mathcal{D}_{A_2} = \emptyset$ implies that $\mathbf{D}_{A_2} = \mathbf{D}(A_1, A_2) - \mathbf{D}_{A_1}$ and therefore that

$$A_1 \prec_{\Omega_\varepsilon} A_2 \quad \textit{iff} \quad \mathbf{D}_{A_1} \leq \frac{\mathbf{D}(A_1, A_2)}{2}. \tag{8}$$

$\mathbf{D}(A_1, A_2)$ is independent from Ω. Consequently one can conclude that $\mathbf{P}(A_1 \prec_{\overline{\Omega}} A_2)$ corresponds to the probability of

$$D = \mathbf{D}_{A_1} - \frac{\mathbf{D}(A_1, A_2)}{2}$$

not changing signs.

Without loss of generality and because of the symmetry of the problem, we can assume that the numbering of A_1 and A_2 is chosen such that $A_1 \prec_{\Omega_\varepsilon} A_2$, and therefore $\mathbf{D}_{A_1} \leq \mathbf{D}_{A_2}$ and thus $D < 0$.

D will switch signs if at least $\hat{D} = \frac{\mathbf{D}(A_1,A_2)}{2} - \mathbf{D}_{A_1}$ events in $\mathcal{D}_{A_2}$ are in disagreement with $\overline{\Omega}$ (and if none of those in $\mathcal{D}_{A_1}$ are). Given that our events are following a Bernouli law of parameter ε, the probability of having at least $\hat{D}$ events (and thus D switching signs) is

$$\sum_{i=\hat{D}}^{\mathbf{D}_{A_2}} \binom{\mathbf{D}_{A_2}}{i} \varepsilon^i (1-\varepsilon)^{\mathbf{D}_{A_2}-i} = \sum_{i=\hat{D}}^{\mathbf{D}_{A_2}} \mathcal{B}(\mathbf{D}_{A_2}, i). \tag{9}$$

This probability is conditioned by the fact that all of the events of $\mathcal{D}_{A_1}$ are in agreement with $\overline{\Omega}$. If k events of $\mathcal{D}_{A_1}$ are in disagreement with $\overline{\Omega}$, then the above probability becomes

$$\sum_{i=\hat{D}+k}^{\mathbf{D}_{A_2}} \mathcal{B}(\mathbf{D}_{A_2}, i). \tag{10}$$

Therefore, the global probability covering all cases, of D switching signs is

$$\mathbf{P}_{\text{switch}} = \sum_{k=0}^{\mathbf{D}_{A_1}} \mathcal{B}(\mathbf{D}_{A_1}, k) \sum_{i=\hat{D}+k}^{\mathbf{D}_{A_2}} \mathcal{B}(\mathbf{D}_{A_2}, i). \tag{11}$$

Finally, since we are looking for the probability of the initial ranking remaining unchanged, we obtain that

$$\mathbf{P}\left(A_1 \prec_{\overline{\Omega}} A_2 \,\middle|_{A_1 \prec_{\Omega_\varepsilon} A_2}\right) = 1 - \mathbf{P}_{\text{switch}} \tag{12}$$

Numerical Example Using the data in Table 3, we observe that $A_1 \prec_{\Omega_\varepsilon} A_2$. Furthermore,

$$\mathcal{D}(A_1, A_2) = \{\phi_4, \phi_5, \phi_6, \phi_7\}$$

$$\begin{aligned} \mathcal{A}(A_1, \Omega_\varepsilon) &= \{\phi_1, \phi_2, \phi_3, \phi_7\} & \mathcal{A}(A_2, \Omega_\varepsilon) &= \{\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6\} \\ \mathcal{D}_{A_1} &= \{\phi_7\} & \mathcal{D}_{A_2} &= \{\phi_4, \phi_5, \phi_6\} \end{aligned}$$

The other parameters we observe from the data are: $\mathbf{D}_{A_1} = 1$, $\mathbf{D}_{A_2} = 3$ and $\hat{D} = 1$. We can therefore rewrite Equation 12 as

$$\begin{aligned} \mathbf{P}(A_1 \prec_{\overline{\Omega}} A_2) &= 1 - \sum_{k=0}^{1} \binom{1}{k} \varepsilon^k (1-\varepsilon)^{1-k} \sum_{i=1+k}^{3} \binom{3}{i} \varepsilon^i (1-\varepsilon)^{3-i} \\ &= 1 - \left((1-\varepsilon)\left(3\varepsilon(1-\varepsilon)^2 + 3\varepsilon^2(1-\varepsilon) + \varepsilon^3\right) + \varepsilon\left(3\varepsilon^2(1-\varepsilon) + \varepsilon^3\right)\right) \\ &= 1 - \varepsilon\left(3 - 6\varepsilon + 7\varepsilon^2 - 3\varepsilon^4\right) \end{aligned} \tag{13}$$

We obtain:

- $\varepsilon = 0.5$: $\mathbf{P}(A_1 \prec_{\overline{\Omega}} A_2) = 0.313$
- $\varepsilon = 0.2$: $\mathbf{P}(A_1 \prec_{\overline{\Omega}} A_2) = 0.589$
- $\varepsilon = 0.1$: $\mathbf{P}(A_1 \prec_{\overline{\Omega}} A_2) = 0.753$
- $\varepsilon = 0$: as expected $\mathbf{P}(A_1 \prec_{\overline{\Omega}} A_2) = 1$

Thus confirming the previously obtained Monte-Carlo estimates.

6 Extension to Multiple Interpretations and Confidence Levels

Until now, we have been considering two algorithms expressing boolean values for a single interpretation. We are still considering two algorithms but expressing confidence values for multiple possible interpretations. In order to handle this case, we are going to use the Kullback-Leibler divergence [4]. The Kullback-Leibler divergence is a dissimilarity measure between two probability distributions P and Q, where P represents a series of observations, or a precisely computed probability distribution, and Q a model or an approximation of P.

Definition 7 (Kullback-Leibler Divergence) *Let P and Q be two probability distributions. The Kullback-Leibler distribution of Q with respect to P is defined by*

$$D_{KL}(P||Q) = \sum_i P(i)\, ln\left(\frac{P(i)}{Q(i)}\right)$$

Note: $D_{KL}(P||Q) = D_{KL}(Q||P) = 0$ *iff* $P = Q$.

6.1 Application to Ranking Evaluation

In order to apply Kullback-Leibler to our case, we need probability distributions. We therefore need to reformulate our problem, and restrict it to some specific cases.

Hypotheses

1. We are still considering two algorithms only.
2. Multiple interpretations are possible (*i.e.* $q \geq 1$).
3. Ground Truth attributes only one interpretation to each data element.
4. Algorithms return a confidence value in [0..1] per possible interpretation for each data element.

We therefore need to redefine the formal concept of an algorithm initially given in Definition 2 as follows:

Definition 8 (Algorithm) *An algorithm A is a function associating a confidence value for one or multiple interpretations to a given data element ϕ.*

$$\begin{aligned} A : \Phi &\rightarrow [0..1]^q \\ \phi &\mapsto (a_1, ..., a_q) \\ &\text{with } \sum_{l=1}^{q} a_l = 1 \end{aligned}$$

As such, the interpretation confidence of a given data element can be assimilated to a probability distribution.

For a given ground truth Ω and a data element ϕ_k we obtain the following Kullback-Leibler distribution:

$$D_{KL}\left(A_i\left(\phi_k\right)||\Omega\right) = \sum_{l=1}^{q} a_{kl}^{j}\, ln\left(\frac{a_{kl}^{j}}{\Omega(\phi_k, i_l)}\right)$$

Application Let A_1 and A_2 be two algorithms to compare. Let $\overline{\Omega}$ be an ideal ground truth of which we have no precise knowledge. Let Ω_ε be a known ground truth, differing from $\overline{\Omega}$ by ε.

First, we establish the ranking between A_1 and A_2 by using the sum of the Kullback-Leibler divergence for all data elements.

$$\mathbf{D}\left(A_i, \Omega_\varepsilon\right) = \sum_{k=1}^{p} D_{KL}\left(A_i\left(\phi_k\right)||\Omega_\varepsilon\right)$$

We can then apply the same definitions and techniques as in the previous sections. $A_1 \prec_{\Omega_\varepsilon} A_2$ *iff* $\mathbf{D}\left(A_1, \Omega_\varepsilon\right) >= \mathbf{D}\left(A_2, \Omega_\varepsilon\right)$ or, in other terms, *iff*

$$\sum_{k=1}^{p} D_{KL}(A_1(\phi_k)||\Omega_\varepsilon) \geq \sum_{k=1}^{p} D_{KL}(A_2(\phi_k)||\Omega_\varepsilon)$$

$\mathbf{P}\left(A_1 \prec_\Omega A_2 \left|_{A_1 \prec_{\Omega_\varepsilon} A_2}\right.\right)$ can now be computed following the same technique as described previously, by replacing the formal divergence formulae with a Monte-Carlo simulation.

	A_1				A_2				Ω_ε			
	i_1	i_2	i_3	i_4	i_1	i_2	i_3	i_4	i_1	i_2	i_3	i_4
ϕ_1	0.3	0.1	0.4	0	0.2	0.6	0.2	0	1	0	0	1
ϕ_2	0.5	0.2	0.1	0.2	0	0.1	0.8	0.1	0	0	1	0
ϕ_3	0.6	0.2	0.1	0.1	0.5	0.3	0.1	0.1	1	0	0	0
ϕ_4	0.4	0.2	0.4	0	0.4	0.3	0.2	0.1	0	1	0	0
ϕ_5	0.1	0.8	0.1	0	0.9	0.1	0	0	1	0	0	0
ϕ_5	0.6	0.3	0	0.1	0.2	0.2	0.2	0.4	0	0	0	1

Table 4. Numerical example for Kullback-Leibler divergence-based ranking evaluation

Numerical Example Using the data in Table 4, with $N = 10^5$ and a 95% confidence we obtain:

- $\varepsilon = 0.5$: $\mathbf{P}(A_1 \prec_\Omega A_2) = 0.6663 \pm 0.0029$
- $\varepsilon = 0.4$: $\mathbf{P}(A_1 \prec_\Omega A_2) = 0.7353 \pm 0.0027$
- $\varepsilon = 0.2$: $\mathbf{P}(A_1 \prec_\Omega A_2) = 0.8665 \pm 0.0021$

7 Conclusion and Perspectives

In this paper we have explored various methods for evaluating algorithm performances with respect to an unreliable ground truth, by expressing the probability that the observed ranking be modified given an error boundary estimate on the ground truth quality.

Our models are able to express the probability of the ranking between two algorithms to flip in presence of a given estimated uncertainty on the ground truth and in the absence of any knowledge of the absolute, untainted ground truth. We have expressed this probability formally in the case of binary interpretation algorithms, and validated it with a Monte-Carlo simulation. We have also formalised the possibility of using Kullback-Leibler divergence in the case of non-binary interpretation algorithms when the interpretations are associated with probability values.

The current limitation of our models is that they yet need to be extended to the ranking of multiple algorithms ($n > 2$) on the one hand, and that the array of possible interpretations be expanded beyond simple binary or probabilistic interpretations, on the other. These theoretical results (although formally proven) also need to be experimentally assessed on real data [6]. Unfortunately, most current published benchmark results only publish precision and recall curves and values, while our methods require access to the complete experimental result set.

Further work will also focus on extending the current conclusions and techniques to simple precision and recall curves.

References

1. Al-Khaffaf, H., Talib, A., Osman, M., Wong, P.: Grec'09 arc segmentation contest: Performance evaluation on old documents. In: Ogier, J.M., Liu, W., Lladós, J. (eds.) Graphics Recognition. Achievements, Challenges, and Evolution, Lecture Notes in Computer Science, vol. 6020, pp. 251–259. Springer Berlin Heidelberg (2010), http://dx.doi.org/10.1007/978-3-642-13728-0_23
2. Al-Khaffaf, H., Talib, A., Osman, M.: Final report of grec'11 arc segmentation contest: Performance evaluation on multi-resolution scanned documents. In: Kwon, Y.B., Ogier, J.M. (eds.) Graphics Recognition. New Trends and Challenges, Lecture Notes in Computer Science, vol. 7423, pp. 187–197. Springer Berlin Heidelberg (2013), http://dx.doi.org/10.1007/978-3-642-36824-0_18
3. Bukhari, S., Al-Khaffaf, H., Shafait, F., Osman, M., Talib, A., Breuel, T.: Final report of grec'13 arc and line segmentation contest. In: Lamiroy, B., Ogier, J.M. (eds.) Graphics Recognition. Current Trends and Challenges, Lecture Notes in Computer Science, vol. 8746, pp. 234–239. Springer Berlin Heidelberg (2014), http://dx.doi.org/10.1007/978-3-662-44854-0_18
4. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Statist. 22(1), 79–86 (03 1951), http://dx.doi.org/10.1214/aoms/1177729694
5. Lamiroy, B.: Interpretation, Evaluation and the Semantic Gap ... What if we Were on a Side-Track? In: Lamiroy, B., Ogier, J.M. (eds.) 10th IAPR International Workshop on Graphics Recognition, GREC 2013. vol. 8746, pp. 213–226. Springer, Bethlehem, PA, United States (Aug 2013), https://hal.inria.fr/hal-01057362

6. Lamiroy, B., Sun, T.: Computing Precision and Recall with Missing or Uncertain Ground Truth. In: Kwon, Y.B., Ogier, J.M. (eds.) Graphics Recognition. New Trends and Challenges. 9th International Workshop, GREC 2011, Seoul, Korea, September 15-16, 2011, Revised Selected Papers, Lecture Notes in Computer Science, vol. 7423, pp. 149–162. Springer (Feb 2013), https://hal.inria.fr/hal-00778188
7. Liu, W.: The third report of the arc segmentation contest. In: Liu, W., Lladós, J. (eds.) Graphics Recognition. Ten Years Review and Future Perspectives, Lecture Notes in Computer Science, vol. 3926, pp. 358–361. Springer Berlin Heidelberg (2006), http://dx.doi.org/10.1007/11767978_32
8. Metropolis, N., Ulam, S.M.: The Monte Carlo Method. Journal of the American Statistical Association 44(247), 335–341 (Sep 1949), http://dx.doi.org/10.2307/2280232