Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200167

Introducing New Measures of Inter- and Intra-Rater Agreement to Assess the Reliability of Medical Ground Truth

Andrea CAMPAGNER [a,b] and Federico CABITZA [b,1]
[a] *IRCCS Istituto Ortopedico Galeazzi, Milano, Italy*
[b] *University of Milano-Bicocca, Milano, Italy*

Abstract. In this paper, we present and discuss two new measures of inter- and intra-rater agreement to assess the reliability of the raters, and hence of their labeling, in multi-rater setings, which are common in the production of ground truth for machine learning models. Our proposal is more conservative of other existing agreement measures, as it considers a more articulated notion of agreement by chance, based on an empirical estimation of the precision (or reliability) of the single raters involved. We discuss the measures in light of a realistic annotation tasks that involved 13 expert radiologists in labeling the MRNet dataset.

Keywords. inter-rater agreement, reliability, Ground Truth, Machine Learning

1. Introduction

Data science research needs valid and reliable data to enable the comprehension of the represented phenomena, and sound statistical inference and prediction about those phenomena, that is the definition of predictive models, e.g., by machine learning methods that can help human decision makers classify and interpret the reality of interest.

In many domains, like medicine, phenomena are increasingly measured by sensors, but are still to a large extent described by human observers, who are supposed to produce data in the endeavor to provide accurate and complete representations, by which to detect effects, trends and differences and build reliable prediction models out of them, what is then called *Ground Truth*.

In the light of the unavoidable fallibility of human observers in providing an ever true representation, in domains characterized by high uncertainty, ambiguity and variability of relevant conditions, like medicine, ground truth datasets are built by combining multiple observations and ratings (i.e., labels), and averaging between them, to associate “the one best” label to each case or object of interest. This is the process by which data scientists can bring together subjective ratings (that is how a single rater sees and interprets a given phenomenon) and create a reliable inter-subjective labelling, which is intended to be the most objective representation of the reality of interest [1].

[1] Corresponding Author: Federico Cabitza, federico.cabitza@unimib.it

Assessing observer variability, that is the extent multiple raters agree (or disagree) in providing a unique interpretation (that is label, or evaluation), is then important to assess the reliability of the ground truth by which to build predictive models from empirical observations. In the literature there are many measures (e.g., Fleiss' Kappa, Cohen's Kappa, Krippendorff's α) and any of them present pros and cons [2]. In particular, to assess inter-rater agreement, the Krippendorff's α is particularly indicated (although seldom used in medicine and by ML scholars) as it is robust with respect to chance effects and missing values.

However, this measures, as the other ones, adopts a naive model of chance, by which to assess the degree by which multiple observers agree with each others beyond the extent they do so by chance. In particular, no measure considers the expertise of the raters involved, nor their confidence in their specific ratings: in short the reliability of their ratings. This brings us to consider an aspect that is seldom considered: how to assess the reliability of the raters involved, beyond self-assessment? This can be evaluated in many ways: we investigated the relationship between this construct and the extent raters agree with themselves in judging the same phenomenon multiple times over time, and hence they do not take guesses. We call this degree, *self-agreement*, while others refer to it with the expression *intra-rater agreement* (e.g., [3]).

In this paper, we will present two new metrics to assess both self-agreement and inter-rater agreement in order to contribute to the existing literature and provide a better tool to assess the reliability of multi-rater labeled datasets.

2. Method

In this Section we first provide a decision-theory based derivation of a measure for *self-agreement*, intended as the probability that a decision-maker gives the same interpretation of the same phenomenon consistently. Next, we use the self-agreement indicator to define a novel chance-adjusted measure of *inter-rater reliability* that we denote as ρ. Let $C = \{0, 1, ..., n-1\}$ be the set of possible class labels, $p = \langle p(0), ..., p(n-1) \rangle$ be the proportion of 0s, 1s, ..., $n-1$s in the actual labelings by multiple raters (we assume that these reflect the real class proportions in the reality of interest).

We assume a two-step decision procedure: first the decision maker flips a biased n+1 faced coin to choose between *random choice* (x) and *peaked choice* ($y_0, ..., y_{n-1}$). Then:

- If the decision maker chose random choice, then she selects one class according to distribution p;
- Otherwise, if she chose peaked choice, she reports one value (which depends on the specific peaked distribution chosen) with probability 1. Notice that while each peaked distribution assigns probability 1 to a single alternative, the specific peaked distribution is chosen according to $(y_0, ..., y_{n-1})$, which allows to encode possible degree of uncertainty of the rater when she does not guess completely at random.

This decision-making model conforms to standard decision-theory where a decision maker is usually modeled as a probabilistic device in particular, our model represents a generalization of the decision model proposed by Krippendorff as a justification for his α metrics [4], in which raters possess different and unrelated probabilities of selecting among the alternatives at random (e.g. due to their different expertise levels).

How can we find x and $y_0,...,y_{n-1}$ given m $o_1,...,o_m$ observations of the decision-maker repeating the same annotation task (i.e. annotating multiple times the same case)? Let $d(0)$, ..., $d(n-1)$ be the observed proportion of 0, 1, ..., n-1 labels, respectively, among $o_1,...,o_m$ (notice that this is a maximum likelihood estimation of the parameters of a nominal distribution). We can find $x, y_0,...,y_{n-1}$ selecting the solution to the following system by maximizing the entropy:

$$\begin{cases} \forall i.d(i) = x * p(i) + y_i \\ x + \sum_i y_i = 1 \\ \forall i.x, y_i \geq 0 \end{cases} \tag{1}$$

We define the *self-agreement* of the observer on label class i as $SA_i = [x * p(i) + y_i]^2 = d(i)^2$ and their overall self-agreement as $SA = \sum_i SA$. Notice that SA_i is the probability that the observer gives the same label i when it is asked to give a label to the same object twice and thus it coincides with the *Gini impurity*.

Starting from our definition of self-agreement, we want to define a measure of inter-rater reliability, i.e. a measure of the extent a set of raters agree in labeling a set of cases or objects. Specifically, we want to define a measure ρ that takes into account the fact that, because that raters do not have perfect self-agreements, mutual agreements may arise due to chance; thus, we want ρ to properly discount this case.

Given two raters i, j and a case x we can find the probability

$$P((i,j) \text{ genuinely agree on x} | (i,j) \text{ agree on x}) \tag{2}$$

by a direct application of *Bayes' rule*. Denoting as $y^i_k(x)$ the value of y_k for rater i on case x, we note that this represents the probability that rater i *genuinely* asserts label i on case x. Denoting as $c_i(x)$ the labeling provided by rater i on case x and supposing that $c_i(x) = c_j(x)$, we can see that, assuming that the raters i, j annotate case x independently:

$$P((i,j) \text{ genuinely agree on x}) = y^i_{c_i(x)}(x) * y^j_{c_j(x)}(x) \tag{3}$$

and

$$P((i,j) \text{ agree on x}) = P((i,j) \text{ genuinely agree on x}) + P((i,j) \text{ agree by chance on x}) =$$
$$y^i_{c_i(x)} * y^j_{c_j(x)} + p[c_i(x)] * x^i * y^j_{c_j(x)} + p[c_j(x)] * y^i_{c_i(x)} * x^j + p[c_i(x)] * p[c_j(x)] * x^i * x^j \tag{4}$$

Then, denoting the ratio of the probabilities in Equations 3, 4 as *Genuine Agreement Effect* (GAE), we can express the ρ measure as:

$$\rho = \frac{1}{|U|} \sum_x \frac{1}{\binom{|D|}{2}} \sum_{i \neq j \in D} \delta(c_i(x), c_j(x)) * GAE(i,j) \tag{5}$$

where $\delta(x,y) = \begin{cases} 1 & x = y \\ 0 & \text{otherwise} \end{cases}$, U is the set of cases and D the set of raters. Notice that when the sample size to estimate the values of the y^i_k factors is small, the estimate obtained via the exact methods is not robust when computing the value of ρ, especially if the value were computed on only a small set of cases and then propagated to the remaining ones (e.g. by averaging). In these cases a more robust but approximate estimate can be obtained by considering ignoring the specific probability of getting a particular peaked choice distribution and only considering $\sum_k y^i_k$ as a measure of the ability of rater i of providing self-consistent annotations.

3. Results

In order to test our proposed measures we considered a dataset obtained by a realistic esperiment of knee *Magnetic Resonance Imaging* (MRI) annotation. Specifically, we asked 13 radiologists from the IRCCS Orthopedic Institute Galeazzi of Milan (Italy) to annotate 417 MRIs from the well-known MRNet dataset [2]: in particular, the doctors were asked to assess, for each of these images, the presence of abnormalities (thus, the considered problem was a binary classification setting, i.e. $C = \{0,1\}$). Furthermore, the doctors were also asked to assess the complexity of each case and the confidence in their annotations, on a 4- and 5-value scale, respectively.

In order to apply our measure, we first had to estimate the self-agreement of each of the raters. In order to do so, we first chose two cases randomly among the ones considered of medium to high difficulty (between 2 and 3 in a 1 to 4 scale) by one of the most experienced radiologist in the panel of experts involved. Then, we inserted these two cases multiple times in the dataset of cases to be labelled, replicating them for 6 times each: in so doing, the dataset had 10 more cases. These replicated images were placed randomly in an annotation sequence (encompassing more than 200 cases) to make more difficult for the radiologists to understand that they had already examined and assessed those cases. Considering the spread, in terms of interquartile range (IQR), in the complexity rating and the varying confidence in the interpretation of these identical cases, we can conjecture that the replicated cases were likely considered different cases by all of the radiologists (First case: complexity IQR = 0.63, confidence IQR = 0.5; Second group: complexity IQR = 0.5, confidence IQR = 0.4).

In regard to the first case, we obtained an average self-agreement of 0.58 ± 0.03 (95% confidence interval, $min = 0.5$, $max = 0.72$) For the second group of images, we obtained an average self-agreement of 0.51 ± 0.01 ($min = 0.50$, $max = 0.56$): notice that these values are close to the minimal value for *SA* which, for the binary case, is 0.50. Averaging between the two groups the average self-agreement was 0.54 ± 0.02 ($min = 0.5$, $max = 0.64$).

In the computation of the ρ, for each agent i, we considered the values $y^i_0 + y^i_1$ estimated on the whole group of 12 repeated images. We obtained a value of $\rho = 0.46$, while for the same dataset we obtained a value of Fleiss' $k = 0.63$ and Krippendorff's $\alpha = 0.63$.

4. Discussion

The above results show how our measure is much more conservative than the others proposed to assess inter-rater agreement. The small difference between the latter measures

[2] https://stanfordmlgroup.github.io/competitions/mrnet/

can be explained by the fact that there was a near perfect class balance: indeed both k and α consider the class balance in order to model chance effects. Furthermore, the large difference between the ρ and the values of the other metrics can be explained observing that the observed self-agreement were indeed very low and the same holds for the obtained y_k^i value (in most cases the value of x^i was near to 0.5). In fact, to obtain a value of $\rho \sim 0.63$ on the dataset, an average GAE around 0.75 would be needed. This shows that, differently from α (and k), the ρ takes into account a model of the rating reliability of the raters involved, and hence it yields a more realistic measure of their agreement: for example, if all raters exhibited a perfect self-agreement (thus $GAE = 1$) we would obtain a value $\rho = 0.82$ while the values for α and k would not change.

Moreover, one could wonder what threshold should be set to assess whether the agreement is sufficient to consider the data reliable, that is what the so-called *smallest acceptable reliability* is [4]. Unfortunately, any proposal of such a threshold would be laden with some extent of arbitrariness. Krippendorff suggests to "not accept data with reliabilities whose confidence interval reaches below the smallest acceptable reliability [...], for example, of .8 00, but no less than .667" [4](p. 242). An α or an ρ below 0.667 would mean that only two thirds of the data are labelled to a degree better than chance". This recommendation challenges a much more popular way to interpret agreement scores since the 1970s by [5], that is much more indulgent (a score of .21 is considered an indicator of *fair* agreement; .41 *moderate*; .61 *substantial*).

As said above, differently from α, the ρ takes into account a model of the rating reliability, and hence it yields a more realistic measure of their agreement. In this first formulation, we have considered self-agreement as a proxy of the rater reliability. Since measuring self-agreement with surreptitious repetitions can be intricate, Formula 5 can alternatively be integrated with a measure of the raters' confidence in the interpretations given (the higher the confidence on a rating, the higher the reliability of that rating as the rater is stating they are not taking a guess). Although this sounds reasonable, yet we did not find confidence to be highly correlated with self-agreement ($correlation = 0.22$, $p - value = 0.46$).

In this regard, however, we should emphasize that low ρ or SA scores should not be used as proxy of rating skills, or to judge how good the raters are: in the MRNet study, the 13 radiologists achieved a remarkable average performance of .82 (min: .78, max: .86) in re-annotating the original, low-res dataset on standard monitors and with no incentives. Rather, what low ρ or SA scores in ground truthing by experts indicate is the intrinsic ambiguity and complexity of medical phenomena; the over-ambition to pinpoint them with clear-cut labels; and the reckless risk to delegate judgment or advice to classifying machines that interpolate those labels as a way to resolve uncertainty.

References

[1] L. Mari. A quest for the definition of measurement. *Measurement*, 46(8):2889–2895, 2013.

[2] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[3] M. E Gianinazzi and et al. Intra-rater and inter-rater reliability of a medical record abstraction study on transition of care after childhood cancer. *PloS one*, 10(5):e0124290, 2015.

[4] K. Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[5] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.