A comparative study of benchmarking procedures for interrater and intrarater agreement studies

Valutazione comparativa di procedure di benchmarking per l'analisi dell'accordo inter e intra valutatore

Amalia Vanacore[1] and Maria Sole Pellegrino[2]

Abstract Decision making processes typically rely on subjective evaluations provided by human raters. In the absence of a gold standard against which check evaluation trueness, the magnitude of inter/intra-rater agreement coefficients is commonly interpreted as a measure of the rater's evaluative performance. In this study some benchmarking procedures for characterizing the extent of agreement are discussed and compared via a Monte Carlo simulation.

Abstract *In numerosi contesti, le decisioni strategiche sono affidate a valutazioni soggettive, fornite da valutatori umani, per le quali non esiste un gold standard che permetta di valutarne la veridicita'. L'affidabilita' del valutatore viene quindi spesso misurata in termini di precisione attraverso coefficienti di accordo inter- e intra-valutatore, che risultanto utili se interpretabili. Nel lavoro proponiamo uno studio Monte Carlo per analizzare e confrontare le prestazioni di alcune procedure di benchmarking.*

Key words: rater agreement, kappa-type coefficient, benchmarking procedures, Monte Carlo simulation

1 Introduction

Agreement coefficients are widely adopted for assessing the precision of subjective evaluations provided by human raters to support strategic and operational decisions in several fields (e.g. manufacturing and service industries, food, healthcare and risk management). Specifically, the agreement between the evaluations provided on the same sample of items by two or more raters (i.e. inter-rater agreement) or by

[1] Dept. of Industrial Engineering, University of Naples ”Federico II”, p.le Tecchio 80, 80125 Naples; email: amalia.vanacore@unina.it

[2] Dept. of Industrial Engineering, University of Naples ”Federico II”, p.le Tecchio 80, 80125 Naples; e-mail: mariasole.pellegrino@unina.it

the same rater in two or more occasions (i.e. intra-rater agreement) is commonly measured using a kappa-type agreement coefficient.

In order to qualify the extent of agreement as good or poor the computed coefficient is compared against an arbitrary benchmark scale. However, the magnitude of an agreement coefficient may strongly depend on some experimental factors such as number of rated items, rating scale dimension, trait prevalence and marginal probabilities [13, 9]. Thus, interpretation based on the straightforward benchmarking should be treated with caution especially for comparison across studies when experimental conditions are not the same.

A proper characterization of the extent of rater agreement should rely upon a benchmarking procedure that allows to identify a suitable neighborhood of the true value of rater agreement by taking into account sampling uncertainty. The most simple and intuitive way to accomplish this task is by building a confidence interval of the agreement coefficient and comparing its lower bound against an adopted benchmark scale. A different approach is the one recently proposed by Gwet [9] which, under the assumption of asymptotically normal distribution, evaluates the likelihood that the estimated agreement coefficient belongs to each benchmark category.

The above benchmarking approaches will be in the following discussed and their performances will be evaluated and compared in terms of weighted misclassification rate via a Monte Carlo simulation study.

The remainder of the paper is organized as follows: in Section 2 two well-known paradox-resistant kappa-type agreement coefficients are discussed; the commonly adopted benchmark scales and some characterization procedures based on parametric and non-parametric approaches to benchmarking are presented and discussed in Section 3; in Section 4 the simulation design is described and the main simulation results are fully discussed; finally, conclusions are summarized in Section 5.

2 Paradox-resistant agreement coefficients

The kappa-type agreement coefficients are rescaled measures of the observed agreement corrected with the probability of agreement expected by chance. The most common kappa-type coefficient is that proposed by Cohen [5]. Despite its popularity, it is affected by two paradoxes [4]: the degree to which raters disagree (bias problem) and the marginal distribution of the evaluations independently provided by each rater (prevalence problem). A solution to face the above paradoxes is to adopt the uniform distribution for chance measurements, which given a certain rating scale can be defended as representing the maximally non-informative measurement system [6].

Specifically, let n be the number of items rated by two raters on an ordinal k-point rating scale (with $k > 2$), n_{ij} the number of items classified into i^{th} category by the first rater and into j^{th} category by the second rater and w_{ij} the corresponding symmetrical weight, introduced in order to consider that, on an ordinal rating scale, disagreement on two distant categories is more serious than disagreement on neighboring categories. The weighted version of the uniform kappa, often referred to as Brennan-Prediger coefficient [9], is formulated as:

$$\widehat{BP}_w = \frac{p_{a_w} - p_{a|c}^{BP_w}}{1 - p_{a|c}^{BP_w}} \tag{1}$$

where p_{a_w}, the weighted observed proportion of agreement, and $p_{a|c}^{BP_w}$, the weighted proportion of agreement expected under the assumption of uniform chance measurements, are respectively given by:

$$p_{a_w} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} \frac{n_{ij}}{n}; \quad p_{a|c}^{BP_w} = \frac{T_w}{k^2} \tag{2}$$

being T_w the sum over all weight values w_{ij}.

Another well-known paradox-resistant agreement coefficient alternative to Cohen's Kappa is the AC_1 coefficient proposed by Gwet [8], whose weighted version AC_2 [9] is formulated as:

$$\widehat{AC}_2 = \frac{p_{a_w} - p_{a|c}^{AC_2}}{1 - p_{a|c}^{AC_2}} \tag{3}$$

where the probability of chance agreement $p_{a|c}^{AC_2}$ is given by:

$$p_{a|c}^{AC_2} = \frac{T_w}{k(k-1)} \cdot \sum_{i=1}^{k} \pi_i (1 - \pi_i) \tag{4}$$

Specifically, $p_{a|c}^{AC_2}$ is defined as the probability of the simultaneous occurrence of two events, one rater provides random rating (R) and the two raters agree (G):

$$p_{a|c}^{AC_2} = P(G \cap R) = P(G|R) \cdot P(R) \tag{5}$$

where $P(G|R) = T_w/k^2$ and $P(R)$ is approximated with a normalized measure of randomness defined by the ratio of the observed variance to the variance expected under the assumption of totally random ratings:

$$P(R) = \frac{\sum_{i=1}^{k} p_i (1 - p_i)}{(k-1)/k} \tag{6}$$

with p_i denoting the propensity that a rater assigns score i to an item which is estimated by $p_i = (n_{i\cdot} + n_{\cdot i})/2n$ being $n_{i\cdot}$ (resp. $n_{\cdot i}$) the total number of items classified into i^{th} category by the first (resp. second) rater.

3 Benchmarking procedures for characterizing the extent of agreement

After computing an agreement coefficient, a common question is ”how good is the extent of agreement?” As a general rule kappa values greater than 0.6 are generally considered acceptable [10]. In order to provide an aid to qualify the magnitude of kappa-type coefficients, a number of benchmark scales have been proposed mainly in social and medical sciences over the years. The scale proposed by Landis and Koch [11] is by far the most widely adopted benchmark scale; it consists of six ranges of values corresponding to as many categories of agreement: poor, slight, fair, moderate, substantial and almost perfect agreement for coefficient values ranging between -1 and 0, 0 and 0.2, 0.21 and 0.4, 0.41 and 0.6, 0.61 and 0.8 and 0.81 and 1.0, respectively. This scale was then simplified by Fleiss [7] and Altman [1], with three and five ranges, respectively, and by Shrout [12] who collapsed the first three ranges of values into two agreement categories.

Despite its popularity, the straightforward benchmarking can be misleading because it does not associate the interpretation of the extent of agreement with a degree of uncertainty and it does not allow to compare the extent of agreement across different studies, unless they are carried out under the same experimental conditions. In order to have a fair characterization of the extent of rater agreement, it is necessary to associate a degree of uncertainty to the interpretation of the coefficient.

Under asymptotic conditions, the magnitude of the kappa type coefficient can be related to the notion of extent of agreement by benchmarking the lower bound of its asymptotic $(1-2\alpha)\%$ confidence interval (CI). Recently, Gwet [9] proposed a parametric benchmarking procedure based on Interval Membership Probability (IMP) that is the probability that the coefficient falls into each benchmark category.

Under non-asymptotic conditions, two non-parametric CIs based on bootstrap resampling are the percentile (p) CI and, for severely skewed distribution, the Bias-Corrected and Accelerated (BCa) CI [2]. Being free from distributional assumptions, the benchmarking procedure based on bootstrap CIs fits also the cases of moderate and small sample sizes.

4 Simulation study

The above-discussed benchmarking procedures have been applied to characterize the extent of both BP_w and AC_2 across 72 different settings. Their statistical properties have been investigated via a Monte Carlo simulation study developed considering two raters classifying $n = 10, 30, 50, 100$ items into one of $k = 2, 5, 7$ possible ordinal rating categories. The data have been simulated by sampling $r = 2000$ Monte Carlo data sets from a multinomial distribution with parameters n and $\mathbf{p} = (\pi_{11}, \dots, \pi_{ij}, \dots, \pi_{ik})$; the π_{ij} values have been set so as to obtain six true population values of agreement (viz. 0.4, 0.5, 0.6, 0.7, 0.8, 0.9), assuming a linear weighting scheme [3].

The performances of the benchmarking procedures under comparison have been evaluated in terms of weighted misclassification rate (hereafter, $\mathbf{M}_{\mathrm{w}}$). Specifically, let $\{X_h; r\}$ be a Monte Carlo data set containing r benchmarks X_h obtained for a population value taken as reference for a specific agreement category ω. $\mathbf{M}_{\mathrm{w}}$ is evaluated as the weighted proportion of misclassified X_h:

$$\mathbf{M}_{\mathrm{w}} = \frac{1}{r} \sum_{\omega=1,\Omega} w_{\omega\omega'} \cdot I\left[X_{h|\omega} \in \omega'\right]; \quad \omega' \neq \omega \tag{7}$$

where $I[\cdot]$ is an indicator taking value 1 if the argument is true and 0 otherwise and $w_{\omega\omega'}$ is a linear misclassification weight adopted to account that on an ordinal benchmarking scale some misclassifications are more serious than others. The best and worst values of $\mathbf{M}_{\mathrm{w}}$ obtained across the analysed benchmarking procedures for BP_w and AC_2 are reported in Table 1 for each combination of n and k values. Specifically, while the benchmarking procedure based on bootstrap CIs are suitable for all the analysed sample sizes, the parametric procedures work only under asymptotic conditions being thus applied only to large samples of $n \geq 50$; therefore the parametric and non-parametric procedures are compared each other only for $n \geq 50$.

Table 1 Best and worst $\mathbf{M}_{\mathrm{w}}$ across the four benchmarking procedures (Standard: Parametric CI; Underlined: IMP; *Italics*: p CI; **Bold**: BCa CI) for BP_w and AC_2 for different n and k values

(a) Best $\mathbf{M}_{\mathrm{w}}$ for BP_w

	$n=10$	$n=30$	$n=50$	$n=100$
$k=2$	*0.102*	**0.096**	*0.068*	0.049
$k=5$	**0.123**	*0.081*	0.056	0.034
$k=7$	**0.087**	*0.066*	0.048	0.027

(b) Worst $\mathbf{M}_{\mathrm{w}}$ for BP_w

	$n=10$	$n=30$	$n=50$	$n=100$
$k=2$	**0.160**	*0.118*	0.080	*0.058*
$k=5$	*0.131*	**0.088**	0.072	0.051
$k=7$	*0.089*	0.072	0.060	0.044

(c) Best $\mathbf{M}_{\mathrm{w}}$ for AC_2

	$n=10$	$n=30$	$n=50$	$n=100$
$k=2$	*0.159*	*0.098*	*0.072*	0.046
$k=5$	*0.111*	*0.073*	0.051	0.030
$k=7$	*0.085*	**0.031**	0.044	*0.026*

(d) Worst $\mathbf{M}_{\mathrm{w}}$ for AC_2

	$n=10$	$n=30$	$n=50$	$n=100$
$k=2$	**0.193**	**0.099**	**0.084**	**0.055**
$k=5$	**0.130**	**0.092**	**0.066**	0.046
$k=7$	**0.092**	*0.058*	**0.056**	0.042

For small and moderate samples (i.e. $n \leq 30$), $\mathbf{M}_{\mathrm{w}}$ slightly differs across non-parametric benchmarking procedures and agreement coefficients: specifically, the highest difference in $\mathbf{M}_{\mathrm{w}}$ is 9%, observed for $n = 10$ and $k = 2$. Moreover, for increasing sample sizes, $\mathbf{M}_{\mathrm{w}}$ becomes quite indistinguishable across procedures and coefficients with a difference always no more than 2%. It is worthwhile to pinpoint that the differences in $\mathbf{M}_{\mathrm{w}}$ across non-parametric benchmarking procedures and agreement coefficients get smaller as n increases because of the decreasing skewness in the distributions of the coefficients: if the distribution is symmetric, the BCa and p CIs agree.

5 Conclusions

The results of the Monte Carlo simulation suggest that for small samples the non-parametric benchmarking procedures based on bootstrap resampling have satisfactory and comparable properties in terms of weighted misclassification rate. Moreover, with $n \geq 30$ the performances of the procedures based on bootstrap CIs differ from each other at most for 2%, therefore benchmarking the lower bound of the percentile bootstrap confidence interval could be suggested — because of its less computational burden — for characterizing the extent of rater agreement, both for BP_w and AC_2. For large samples, the performances are indistinguishable across all benchmarking procedures, thus benchmarking the lower bound of the parametric confidence interval would be preferred being the easiest method to implement.

References

1. Altman, D.G.: Practical statistics for medical research. CRC press (1990)
2. Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med **19**(9), 1141–1164 (2000)
3. Cicchetti, D.V., Allison, T.: A new procedure for assessing reliability of scoring EEG sleep recordings. Am J EEG Technol **11**(3), 101–110 (1971)
4. Cicchetti, D.V., Feinstein, A.R.: High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. **43**(6), 551–558 (1990)
5. Cohen, J.: A coefficient of agreement for nominal scales. Educ Psychol Meas **20**(1), 37–46 (1960)
6. De Mast, J., Van Wieringen, W.N.: Measurement system analysis for categorical measurements: agreement and kappa-type indices. J Qual Technol **39**(3), 191–202 (2007)
7. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol Bull **76**(5), 378–382 (1971)
8. Gwet, K.L.: Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol **61**(1), 29–48 (2008)
9. Gwet, K.L.: Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC (2014)
10. Hartmann, D.P.: Considerations in the choice of interobserver reliability estimates. J. Appl. Behav. Anal. **10**(1), 103–116 (1977)
11. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics pp. 159–174 (1977)
12. Shrout, P.E.: Measurement reliability and agreement in psychiatry. Stat. Methods Med. Res. **7**(3), 301–317 (1998)
13. Thompson, W.D., Walter, S.D.: A reappraisal of the kappa coefficient. J Clin Epidemiol **41**(10), 949–958 (1988)