Measuring and Optimizing the Quality of Crowd-sourced Human Rating

Gary Feng; Lei Chen

November 16, 2015

Introduction

The overarching goal of this project is to explore ways in which we can simultaneously increase the quality of human ratings of multimodal performances while reducing the cost and time required. The approach we take is to use multiple minimally trained human raters, and to rely on the average (or other forms of aggregated) rating as a more reliable and valid measure of performance. Toward this end, we need to evaluate the reliability of individual raters as well as the average rating. We will also need ways to estimate the number of raters needed to achieve a predefined reliability for any single performance. In practice an effective system should be able to identify inconsistent raters as well as performance that generate non-converging responses.

The first part of the paper attemps to identify a metric for reliability for aggregated ratings in the classic testing theory. We start with the intraclass correlation (ICC) family of measures (Bartko, 1966, 1976; Bonett, 2002; Fleiss & Cohen, 1973; Haggard, 1958; Lahey, Downey, & Saal, 1983; McGraw & Wong, 1996; Müller & Büttner, 1994; Shrout & Fleiss, 1979; Swiger, Harvey, Everson, & Gregory, 1964).

- [ ] (Shold we discuss reliability in the generalizability framework, (Dimitrov, 2002)?)

Reliability measures defined in the classic testing theory may not be adequate for issues arise in a crowd-sourced rating application. The second part of the paper therefore outlines unique challenges in the definition and application of the notion of reliability in a rating study where human raters are dynamically allocated to scoring work products in order to maximize the quality and minimize the cost of rating.

Metrics of Quality of Human Scoring

Scoring function and scoring quality

Let us start by defining a scoring function, which is at the center of human rating and automated scoring.

Consider an artifact A (e.g., a multimodal recording of a performance to be evaluated) represented as a multidimensional feature vector $\tilde{\mathbf{A}}$ that is to be assigned a numeric value x_i accoring to a certain criterion ($r_i : \tilde{\mathbf{A}} \mapsto \Re$), that maps $\tilde{\mathbf{A}}$ to a real valued score by a rater i. We further assume that there exists a criterion r (aka a rubric) that defines the "true" mapping ($r : \tilde{\mathbf{A}} \mapsto \Re$), thus $x = r(\tilde{\mathbf{A}})$ is the "true score" of A according to r.

We can further speculate that the mapping occurs in two steps. The first is a mapping from the feature set to a probability distribution of possible scores, and a second step is taken to choose the value with the highest probability. Hence we can redefine the scoring function as:

$$r : \tilde{\mathbf{A}} \mapsto p(x)$$

where $p(x)$ is the likelihood of value x represented as a probability distribution. The rater always chooses x with the greatest likelihood.

The quality of x_i is intuitively the reduction of uncertainty about x when x_i is known. In other words, the posterior distribution of the true score should be narrowed after rater i rated.

$$p(x|x_i)$$

The above is construed differently from the classic testing theory due to the application we will pursue in Section 2.

- The artifact A and its feature set $\tilde{\mathbf{A}}$ is invariant, as the feature set will be extracted algorithmically in automated scoring models. Here we also assume that all raters perceive the features equally, though they would weigh the features differently in scoring.
- Differences in ratings across raters are caused by differential mapping functions r_i. This is a departure from the general notion in the classic testing theory, where it is typically assumed that $x_i = x + e_i$ where e_i is *i.i.d.*. We push the individual differences from scores to the scoring function because an automated scoring function is essentially an algorithmic reliazation of such a scoring function. Furthermore, we anticipate idiosyncratic yet self-consistent scoring functions for individual raters in a crowd-sourced human scoring study. We need to model these rather than simply sweeping them under the rug of error variance.
- We construe the quality of a score in term of how it shed light on the "true" score, not necessarily on how numerically close it is to the true score. We do not assume the numeric value is on something more than a nominal scale. For example, x may represent a label for a statement. In this case, $x_i - x$ is undefined. However, $p(x|x_i)$ always is.

Scoring quality in the classic testing theory

Our immediate concern is that we have conducted a generalizbility study (**???**) where multiple raters rated all artifacts. We need to estimate the quality of the rating. Classic testing theory provides the convenient machinary.

We now review how quality is defined in classic testing theory frameworks. We will come to focus on the ICC family indecies.

Forms of ICC

Numerous forms of ICCs have been identified in the literature (McGraw & Wong, 1996; Shrout & Fleiss, 1979). They can be derived from an ANOVA framework, depending on assumptions about the rating situation.

Let's start with a model presented by the classic McGraw and Wong (1996) paper. We shall focus on the derivation of the ICC(c, k) model.

More generally, the following classes of ICCs are often used in practice.

ICC(1,1)

A case where different raters rated each item, or an one-way ANOVA model.

ICC(2,1)

This correspond to a two-way ANOVA with random effects for both the rater and subjects. The rationale for designating raters as random effect, however, may be several (Hancock & Mueller, 2010, p. 151). One possibility is when the decisions of the study will be based on absolute scores (e.g., when there is an absolute cut-off score). Another common case is when the investigator wishes to use the fully-crossed G study results to estimate future D study where raters may not be fully-crossed; i.e., when we will in the future use a different rater to rate each subject.

ICC(3,1)

ICC(3,1) corresponds to a mixed-effect two-way ANOVA with random effect for subjects but fixed effects for the rater.

ICC with k raters

When the intended measure of scoring quality is based on the average of all k raters, we arrive at the following models.

Choosing a reliabilty measure

In the classic paper Shrout and Fleiss (1979) couched measurement selection in the ANOVA framework. In our case, where the primary interest is in the reliability of the mean score, we are in effect conducting what Shrout and Fleiss (1979, p. 426) called "a substantive study (D study)." The reliability of the mean rating is always greater than that of individual raters (Lord, Novick, & Birnbaum, 1968).

Shrout and Fleiss recommended that the number of observations (m) used to form the mean should be determined by a pilot reliability study (G study); see the next section. Alternatively this may be decided on "substantive grouns," meaning by practical considerations. Once a minimal number of individual raters is established, the reliability of the average rating of the m raters can be estimated using the Spearman-Brown forula and the ICC model for single rater reliability index, i.e., ICC(1,1), ICC(2,1) or ICC(3,1). **Huh? This doesn't make much sense**

Numerous authors have provided decision guidelines based on the Shrout-Fleiss model (e.g., Hancock & Mueller, 2010, p. 151). McGraw and Wong (1996) extended the Shrout and Fleiss family of ICCs and provided a decision tree for selecting the ICC measure for different use cases. According to the M-W framework (McGraw & Wong, 1996, p40), the appropriate model for our case, namely a two-way, random effect, average measure, consistency-based index, should be the ICC(c, k)

Muller and Buttner (1994) provided a decision tree to guide the selection of a reliability metric for a particular rating study. According to the decision tree, if we assume that raters are randomly selected, each rater rates each subject, and measurements are **not** exchanable, the correct measurements are Model B (k, P), Lin (2, P), or Kappa (2, NP). **Check to make sure they are ICC(2,k)**

Determining the number of raters

The number of raters required to achieve a certain level of reliability is discucussed in (Shrout & Fleiss, 1979). Related, (Bonett, 2002) gave an approximation for deteriming the sample size requirements for estimating intraclass correlations with desired precision.

Note, however, that these discussion assume that the number of raters will be fixed for different artifacts, at least missing values are none sysmatic.

In our application, however, we seek to assign different raters and different numbers of raters to each artifact. The assignment will be non-random. This violates the basic assumptions of the classic testing theory based derivations. Hence the results here are not applicable.

Calculating the ICC and related statistics

In R ICC can be calculated using the function "icc" with the packages psy or irr, or via the function "ICC" in the package psych.

Optimizing Scoring Quality for Crowd-sourced Scoring

Our goal in the second section is to design a scoring model where raters will be dynamically assigned to artifacts in order to maximize the quality of the scoring while minimizing a cost function for scoring. Results from the classic testing theory do not tell us how to dynamically allocate scoring resources to optimize scoring quality. Nevertheless they provide some basic parameters to start.

Defining a cost function

The cost of a rater i to score an artifact A_j is function of

- the time spent to score A_j
- the hourly rate of the rater i, assuming the rate is constant

though additional factor may come into play:

- the availability of i, i.e., the wait time for i to rate A_j
- the opportunity cost of i not rating other artifacts

But in general we assume the cost of the rater is independent from the artifact A, or at least conditionally independent after some simple covariates are identified (e.g., the length of video recording).

Quality of rater

Let's attempt to define the quality of a rater. For a specific artifact A_j, the quality is the reduction of uncertainty about the true score distribution given x_i.

Over the population of A the quality can be defined over the joint distribution, which is strict. We could also define the quality based on the marginal distributions, which leaves open the rater-item interaction, where r_ij is different from r_i. We shall deal with the simple case first.

Optimizing for quality

Here is a simple strategy for optimization: Given a pool of m artifacts and n raters each having a scoring function $r_1 \dots r_n$ (or $r_1j \dots r_nj$ for the artifact A_j, where $j = 1 \dots m$), we set a minimal threshold of quality, and minimize the cost function.

References

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, *19*(1), 3–11. Retrieved from http://www.amsciepub.com/doi/abs/10.2466/pr0.1966.19.1.3

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin*, *83*(5), 762. Retrieved from http://psycnet.apa.org/journals/bul/83/5/762/

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine*, *21*(9), 1331–1335. Retrieved from http://www.researchgate.net/profile/Douglas_Bonett/publication/11266478_Sample_size_requirements_for_estimating_intraclass_correlations_with_desired_precision/links/5463d3d50cf2837efdb34670.pdf

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement*, *62*(5), 783–801.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*. Retrieved from http://psycnet.apa.org/psycinfo/1974-04128-001

Haggard, E. A. (1958). Intraclass correlation and the analysis of variance. Retrieved from http://psycnet.apa.org/psycinfo/1959-02517-000

Hancock, G. R., & Mueller, R. O. (2010). *The reviewer's guide to quantitative methods in the social sciences*. Routledge.

Lahey, M. A., Downey, R. G., & Saal, F. E. (1983). Intraclass correlations: There's more there than meets the eye. *Psychological Bulletin*, *93*(3), 586. Retrieved from http://psycnet.apa.org/journals/bul/93/3/586/

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30. Retrieved from http://psycnet.apa.org/journals/met/1/1/30/

Müller, R., & Büttner, P. (1994). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine*, *13*(23), 2465–2476. Retrieved from http://www.en.msc-epidemiologie.med.uni-muenchen.de/download/winter_term_2011_2012/epiresearchdesigns_articles/muellerbuttner02_12.pdf

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420. Retrieved from http://psycnet.apa.org/journals/bul/86/2/420/

Swiger, L. A., Harvey, W. R., Everson, D. O., & Gregory, K. E. (1964). The variance of intraclass correlation involving groups with one observation. *Biometrics*, 818–826. Retrieved from http://www.jstor.org/stable/2528131