Distributional Ground Truth: a Non-Redundant Data Quality Control Method

Maxim Bakaev [0000-0002-1889-0692]

Novosibirsk State Technical University, Pr. K. Marksa 20, Novosibirsk 630073, Russia
bakaev@corp.nstu.ru

Abstract. Despite the growing abundance of data and its growing importance with respect to the booming Machine Learning techniques, certain domains are still choked by limited datasets. Obtaining human-labeled data of proper quality is generally costly, as these activities are often outsourced to low-motivated performers of Crowdsourcing platforms. To validate their outcome, typically such redundant data quality control methods as ground truth and majority consensus are used. In our paper we propose a non-redundant method for prediction of crowdworkers' output quality in web UI labeling tasks, based on homogeneity of distributions assessed with two-sample Kolmogorov-Smirnov test. Using a dataset of about 500 screenshots with over 74,000 UI elements located and classified by 11 trusted labelers and 298 Amazon MTurk crowdworkers, we demonstrate the advantage of our approach over the baseline model based on mean Time-on-Task. Exploring different dataset partitions, we show that with the trusted set size of 17-27% UIs our "distributional ground truth" model can achieve R^2s of over 0.8 and help to obviate the ancillary work effort and expenses.

Keywords: Training Data Quality, Crowdsourcing, Image Labeling, User Interfaces.

1 Introduction

The threshold of data volume that allows AI/ML methods to become effective for most practical tasks is estimated as millions of labeled data samples [1]. They say that data are big and ubiquitous nowadays, but relevant and high-quality data is not so easy or cheap to obtain, particularly if it involves human activities [2, 3]. In most domains, the best training datasets are produced in the process of data labeling, but employing appropriately trained and motivated specialists for this end is generally costly.

Correspondingly, relatively few research groups and IT industry development teams can afford full-time annotators, and data labeling tasks are often outsourced. Typically for any outsourcing, this imposes additional requirements for controlling the **quality** of the outsourced task's outcome. Perhaps not so typical is that the performers of the outsourced data labeling are by and large non-specialists, have low motivation and get rather low wage for their work [4]. The reason is that many ML-related Human Intelligence Tasks (HITs) just require their performer to be a human, and are thus rather tedious, repetitive and trivial.

1.1 Related Work in Data Quality Control

Currently, much of the outsourced data labeling is done via the so-called crowdsourcing (crowdworking) platforms, whose development went hand in hand with the present AI/ML boom: Amazon Mechanical Turk (AMT) (2005), microworkers.com (2009), Yandex.Toloka (2014), Google’s AutoML (2018), etc. The core challenge in crowdsourcing today is obtaining data of appropriate quality [5], and the platforms struggle to support data quality assessment and control [6]. Many meta-tools that emerged lately are aimed specifically on adding such effective capabilities: CDAS, Crowd Truth, iCrowd, DOCS for AMT, and so on [7]. A similar trend is the platforms that specialize in tasks of a particular kind or from a specific domain, e.g. Mighty AI, Hive (.ai), Scale (.ai), etc. for AI in automotive industry [8].

A comprehensive review of quality control methods for crowdsourcing is provided in [6], where the methods are organized into three major groups: individual, group and computation-based. The former two generally imply involvement of trusted humans into assessment of the annotators or of the tasks output, thus suggesting additional overhead in the work effort. Of the crowd data quality control methods, majority/group consensus (MC) and ground truth (GT) are arguably the most widely used, being also supported in most of the platforms [6]. A trendy direction of research within this sub-field is supplementing these two methods for more effective aggregation of the results or allocation of the workers. For instance, in [9] the authors proposed statistical quality estimation based on a two-stage probabilistic generative model for crowdsourcing tasks implying unstructured output. In [10] they studied optimal distribution of training set answers, and it was shown that accuracy of majority voting is highest if the labels in training data follow a uniform distribution. An extension of the GT method based on probability distributions was proposed in [11], where the decision whether to run a second crowd labeling on an image depends upon the “trusted” distribution of labels. In any case, MC and GT imply redundancy (several trusted or candidate workers performing the same task), which has been the mainstream approach in the crowdsourcing quality assessment for already more than a decade [9].

The third group of the quality control methods, the computation-based ones, holds the potential for non-redundant data quality control, which could effectively decrease the share of unnecessary data that essentially goes to waste. The two relevant approaches are outlier analysis that is duly mentioned in [6] and interlaboratory comparisons that is widely used in Metrology [12]. An example of the former applied to crowdsourcing is [13], where they evaluated performance of UIs. The authors did not find statistically significant differences for the data collected in the lab settings and through the AMT platform. This however required resolving certain practical challenges, particularly in outlier detection.

The methodology of the ISO 13528:2015 Statistical methods for use in proficiency testing by interlaboratory comparison standard is mostly built upon the “allowed error” and includes comparison of the measurements to some known distributions [12, ch. 8.4]. Similarly, certain methods in crowdsourcing quality control, some of which are already implemented in the platforms, involve comparing the workers’ outcome or some secondary parameters to their “trusted” values. The foremost example is the

Time-on-Task parameter [6], for which reasonable thresholds are relatively easy to set up. A malicious crowdworker, however, after learning about this and similar behavior-related parameters can adjust their behavior to manipulate them, thus nullifying the quality control. So, we believe that "break-proof" approaches in this field need to be based on non-obvious characteristics of the outcome. In particular, in developing our "Distributional Ground Truth" (DGT) method we propose focusing on statistical distributions in trusted data and measure the candidate data's departure from them.

1.2 The Research Question

In our previous research work [14], we found that crowdworkers' output quality is statistically significantly correlated with the fit of a frequency distribution in their data to a power law, which acted as a sort of "ground truth". The $R^2 = 0.504$ that we obtained was somehow better than the $R^2 = 0.432$ value for the baseline quality control parameter, the Time-on-Task. In the current paper, we introduce a trusted dataset and retrieve a "ground truth" distribution from it, instead of relying on a "common sense" or a domain-dependent distribution.

The evaluation of the DGT method's applicability is done for user interface (UI) screenshot labeling task that is gaining in popularity as computer vision methods are seeing wider application in human-computer interaction (HCI), particularly for UI visual analysis [15], but for which no dedicated quality control methods had been developed, to the best of our knowledge. Meanwhile, HCI does extensively rely on crowdsourcing: e.g. for solving small UI design problems at a large leveraging diversity of microtasking results in CrowdDesign [16], to adapt web layouts to a variety of different screen sizes in CrowdAdapt [17], and so on.

The rest of the paper is organized as follows. In Section 2, we present the method and the involved apparatus, including Kolmogorov-Smirnov test for 2 samples, which is the statistical foundation of the DGT method. Then we describe our experimental study for its evaluation, which involved three sessions: 1) with the 11 trusted labelers, 2) with the 22 verifiers of the trusted labelers' output, and 3) with 298 AMT crowdworkers. In Section 3 we analyze the data, evaluate the DGT method and compare the results to the baselines in data quality control. In the final sections, we discuss the results, note the limitations of our findings, present the contributions and take-aways and outline directions for further research.

2 The Method and the Experimental Study

2.1 The Distributional Ground Truth (DGT) Method

With respect to the classifications provided in [6], the DGT method that we propose falls into such categories as "data quality", "accuracy", "computation-based" and "outlier analysis". The principal idea is testing of distributions' equality (homogeneity), i.e. calculation of statistical distance between the distribution in a trusted dataset and in the assessed dataset (in our particular case, produced by crowdlabelers of uncertain motivation and skill). In this, we do not mean the distributions as probability distributions or generalized functions, but rather as frequency distributions.

The distance measure appropriate for the purposes of the method can be provided by Kolmogorov-Smirnov (KS) nonparametric test for two samples. The test compares the samples' cumulative distributions and computes p-value that depends on the largest discrepancy (distance) between the distributions [18]. Being more powerful than Mann-Whitney's test used to compare the medians of two unpaired groups of data, KS test is sensitive to differences in both location and shape of the distributions. The null hypothesis is that both samples are randomly drawn from the same set of values. Among the assumptions of KS test for two samples are:

- the samples are mutually independent,
- the scale of measurement is at least ordinal,
- the variables are continuous.

Of the two statistics provided by KS test, we are going to rely on p-values, since the distance measures D may have different degrees of freedom and are not directly comparable. Since the values for the KS distribution functions are known and tabulated, the computational complexity of the test is defined by the sorting stage, and thus is not worse than $O(n^2)$, some algorithms even reducing it to O(n) [19]. Another advantage of the KS test is that it can work with relatively low number of values in the two distributions, unlike for testing the fit to power law [20]. Actually, when the samples sizes are close, like we plan to have, increasing one sample may lead to the paradoxical higher bias in the KS test [21]. So, by design, the DGT method is appropriate for application with a limited number of samples typical for crowdsourced HITs, unlike power law that we used in [14].

2.2 The Experimental Evaluation

The objective of the experiment was to explore the effectiveness of the DGT method in crowdsourced data quality control and to estimate the efficient size of the trusted set. The hypothesis is that the DGT method can be used to better explain performance of crowdworkers in UI labeling tasks compared to the baseline or the alternative factors we are about to consider. An important note is that although the trusted labelers and the crowdsourcers had worked with the same material, the study design ensures that the trusted and the testing sets never overlap, so redundancy does not emerge.

The Trusted Set Labeling. The objective of our first experimental session was to obtain the trusted set that could provide the "distributional ground truth" in our study.

Material. The material for the UI labeling was screenshots of homepages of websites belonging to higher educational organizations (universities, colleges, etc.). These were collected by a dedicated Python script crawling through URLs that we acquired from various catalogues (DBPedia, etc.) and then undergoing manual screening (see in [14]). Overall, 497 screenshots were selected based on the following criteria:

- University or college corporate website with reasonably robust functionality;
- Not overly famous university;
- Website content in English and reasonably diverse (i.e. no photos-only websites);
- Reasonable diversity in website designs (colors, page layouts, etc.).

Participants. The trusted labelers were student members of the Novosibirsk State Technical University crowd-intelligence lab, who volunteered to work in the project and provided informed consent. In total, there were 11 of them (6 male, 5 female), with age ranging from 20 to 24 (mean = 20.5, SD = 0.74). All the labelers had normal or corrected to normal vision and reasonable experience with web UIs and IT.

Procedure. To perform the task, the participants used LabelImg (https://github.com/tzutalin/labelImg), a third-party dedicated software tool they were asked to install on their computers. It allows drawing bounding rectangle around an image element, specifying a label for it, and saving the results as XML files (PASCAL VOC format). The screenshots were distributed among the participants near evenly, but no random assignment was performed. The labelers worked independently and on their own computer equipment, and each of them was provided with the identical instructions manual.

Design. We have devised the list of 20 labeling classes for the UI elements. In that, we sought to cover the three major groups of visual objects specific for web UIs: graphical content elements, textual content elements, and interface elements. The names and descriptions of the classes can be found in [7, Table 1].

The output of the labeling was the collection of XML files, each corresponding to its UI. As per Pascal VOC format, for each labeled UI element there was the specification of the bounding box (`xmin-top left, ymin-top left, xmax-bottom right, ymax-bottom right`) and the name of the class. Using dedicated scripts, we derived the following variables for each of the 11 labelers:

- distribution of classes, i.e. the number of labels in each class (both pre-defined and custom);
- mean number of labeled elements per UI: EUI_T.

As the labeling was performed at the participants' convenience, we did not measure the Time-on-Task.

The Labeling Quality Verification. The objective of the second experimental session was to obtain the assessments of the trusted labelers' performance.

Material. The verification was performed for the UIs produced by the trusted labelers. Each labeled UI was represented as the combination of the screenshot file (exactly the same as the trusted labelers used) and the Pascal VOC XML file containing the labeling results. These were rendered together in dedicated web-based software, which would add the verification information to the XML.

Participants. The total number of participants who performed the verification was 20 (10 male, 10 female), and their age ranged from 20 to 22 (mean = 21.1, SD = 0.45). They were next year's students of the Novosibirsk State Technical University crowd-intelligence lab, and none of them had participated in the aforementioned labeling. In a similar fashion, they volunteered to work in the project and provided informed consent. All the participants had normal or corrected to normal vision and reasonable experience with web UIs and IT. They did not report previous experience of working with labeling tools, and they were provided with a specially developed instruction.

Procedure. The labeled UIs were distributed among the 20 verifiers near evenly, but without random assignment. The verification process was performed independently for each UI element in each screenshot, so that the element's labeling could be identified as *correct* or *incorrect*. The reasons for a label to be marked as incorrect were described in the detailed instructions provided to the verifiers and included: too much empty space in the bounding box, cutting neighboring UI elements (except for nesting), incorrect object class, etc. Also, for each labeled UI the verifying participant was asked to provide subjective assessment of the labeling completeness, i.e. if all the visible UI elements were labeled.

The participants worked with a dedicated web-based verification software that we created. Given a set of screenshot image files and corresponding label files in Pascal VOC XML serialization, it allowed to quickly navigate and verify the labeled UIs.

Design. From the binary correct/incorrect data for each UI element recorded by the verification software, we calculated the $\text{Precision}_\text{T}$ for each i-th trusted labeler, as the average precision per the Ni processed screenshots:

$$Precision_{Ti} = \text{avg}_{N_i} \frac{correct}{correct+incorrect} \tag{1}$$

The subjective completeness (SC) ranging from 0 (lowest) to 100 (highest) for each of N_i labeled UIs was similarly averaged for each i-th trusted labeler:

$$SC_i = \text{avg}_{N_i} SC\,/100 \tag{2}$$

The ultimate quality index (Q) reflecting the performance of each i-th trusted labeler was then defined as follows:

$$Q_i = Precision_{Ti} * SC_i \tag{3}$$

The quality index thus incorporated both precision and completeness of the labeling and was subsequently used to order the labelers in the trusted set.

The Crowdsourced Labeling. The objective of this session was to collect the data from crowdworkers, for the subsequent comparison with the trusted set. In order to utilize the distributional ground truth method, we needed enough UIs and UI elements to form distributions of the results for each crowdworker. So, our HIT (Human Intelligence Task) in AMT was designed accordingly, as described below.

Material. The material for the crowdsourced labeling was the same screenshots uploaded to AMT (only the PNG files). From the 497 initial screenshots, 2 were excluded due to the aforementioned technical problems.

The budget allocated for the AMT experimental session was 300 USD, in accordance with our estimation of an average UI labeling task difficulty and the required work effort of 5 minutes.

Procedure (HIT). The labeling HIT was designed using the `Crowd HTML elements` provided by AMT, based on the `crowd-form` and `crowd-bounding-box` widgets, with the screenshot URL as input parameter. HITs could be previewed and skipped by the crowd workers.

Over a time span of 44 days, the labeling and set HITs were available on AMT in 4 batches of 80, 160, 160 and 97 screenshots. Within a batch, workers could submit as many labeling HITs as they wanted. To increase the diversity, however, workers who had successfully labeled 20 or more screenshots in a batch were not allowed to accept labeling HITs in the following batch.

Design. Exactly one label per bounding box and only labels from the list of predefined classes could be selected by the crowd workers. However, the classes were reorganized and their number decreased to 10 (see the list of classes in [14]), due to the following considerations:

- generally lower motivation of the crowdworkers;
- to explore if the distributional ground truth method can work independently of the particular list of classes in the trusted and the testing set.

The obviously malicious contributions would be rejected after our quick visual inspection, and crowdworkers who repeatedly submitted them were excluded from further submissions. All other submissions were approved and received the rewards (on average, 0.26 USD per accepted HIT). From the UI labeling results and the data recorded by AMT, we derived the following variables for each worker:

- distribution of classes, i.e. the number of labels in each class;
- mean number of labeled elements per UI (HIT): EUI_{AMT};
- mean Time-on-Task for each worker: ToT_{AMT};
- $\text{Precision}_{\text{AMT}}$, as the reflection of the worker's performance, was calculated based on the number of accepted and rejected HITs for the worker, in a manner analogous to (1):

$$Precision_{AMT} = \frac{acceptedHITs}{acceptedHITs+rejectedHITs} \tag{4}$$

Participants. The HIT was generally favorably encountered by AMT crowdworkers, with no negative comments or complaints. It total, there were 298 recorded workers, but 20 of them were blocked as malicious. According to the geo information provided by AMT, most of them were from: USA (44.8%), Brazil (15.5%) and India (13.8%).

3 Results

The dataset collected in our study is available at https://figshare.com/s/356b11d5b117014deda2. The code in Python and R is available at https://figshare.com/s/582f5aa0b564e0512f41. Some additional detail on the crowdworkers' budget can be found in [14].

3.1 Descriptive Statistics

The Trusted Set. In total, the labelers processed 495 UIs (2 screenshots had technical problems and were discarded) with 42716 labeled UI elements, of which 39803 (93.2%) belonged to the 20 pre-defined classes. We did some minor adjustments in the erroneous custom classes (e.g., joining *textt* with *text* and *link'* with *link*).

During the verification of the labeling, two more UIs (0.4%) were removed from further analysis due to technical problems with the XML files, so 493 UIs remained. For them the 20 verifiers provided 37574 correct and 4977 incorrect ratings for the labeled UI elements, as well as the SC assessments for 487 UIs (for another 6, SC wasn't specified). In Table 1 we show the statistics for the trusted labelers ordered by the quality index (3) that incorporates both Precision$_T$ and SC. Their order in the trusted set will be considered in the subsequent DGT method application.

Table 1. The trusted labelers' quality of work verified.

Labeler	UIs	EUI$_T$	Precision		SC		Quality
			Mean	SD	Mean	SD	index (Q)
VY	43	87.4	0.928	0.150	95.5	7.0	0.886
SV	44	80.0	0.974	0.056	80.4	12.9	0.783
KK	44	89.3	0.944	0.105	82.5	11.5	0.779
GD	44	121.6	0.899	0.078	84.3	8.1	0.758
PV	44	113.5	0.916	0.180	81.7	17.1	0.748
SMl	43	105.9	0.895	0.078	77.5	11.6	0.694
NE	44	61.6	0.851	0.197	78.3	24.2	0.666
AA	55	89.3	0.890	0.151	73.0	15.1	0.649
PE	43	39.6	0.779	0.147	72.0	17.2	0.561
MA	44	76.0	0.720	0.136	75.1	12.2	0.541
SMr	45	85.1	0.959	0.082	56.0	29.0	0.537
Total/mean	**493**	**86.3**	**0.887**	**0.149**	**77.7**	**18.7**	**0.698**

The correlation between EUI$_T$ and SC was not significant (r_{11} = 0.496, p = 0.121), which might suggest that the "true" number of elements in UIs is variable, even after averaging in reasonably large samples (43-55 UIs). Neither was EUI$_T$ significantly correlated with Precision$_T$ (r_{11} = 0.478, p = 0.137) or Q (r_{11} = 0.461, p = 0.154).

The AMT set. In total, we collected 31676 labeled UI elements for 488 accepted and 754 rejected HITs. The rejection reasons were as follows (one reject could combine several reasons, so they do not sum up to 754):

- ‘incomplete labeling – there are significantly more objects in the screenshot’: 415;
- ‘groups of objects labeled together instead of individually’: 239;
- ‘imprecise bounding boxes’: 207;
- ‘randomly labeled non-existing objects’: 178;
- ‘empty submission’: 139;
- ‘wrong object types labeled’: 79.

The mean $\text{Precision}_{\text{AMT}}$ was 0.442 (SD = 0.475), as opposed to the mean $\text{Precision}_{\text{T}}$ = 0.887 (SD = 0.149) in the trusted set (even though the verification process in the latter was considerably more thorough), which reinforces the need for the crowdsourcing data quality control. The mean EUI_{AMT} was 28.2 (SD = 21.1), i.e. 3 times lower than EUI_{T}. The mean number of UI elements in **accepted** HITs was 58.3, still 1.6 times lower than the respective number in the trusted set. The correlation between the EUI_{AMT} and $\text{Precision}_{\text{AMT}}$ per workers turned out to be highly significant ($r_{298} = 0.751$, $p < 0.001$), unlike in the trusted set ($r_{11} = 0.478$).

The total amount of time spent on the 1242 HITs by all the workers was 665322 s, and on average a worker devoted 635 s (SD = 481 s) to a UI labeling HIT, the correlation between ToT_{AMT} and $\text{Precision}_{\text{AMT}}$ being significant ($r_{298} = 0.449$, $p < 0.001$), but considerably lower than for EUI_{AMT}. The correlation between ToT_{AMT} and EUI_{AMT} was also significant, but not as high as one might expect ($r_{298} = 0.580$, $p < 0.001$). The mean Time-on-Task turned out to be more than twice as long compared to the 5 minutes (300 s) that we estimated when planning the crowdsourcing session budget. Interestingly, 22 workers who didn’t label a single UI element still spent 188416 s on the HITs, which reinforces our concerns about manipulating the Time-on-Task.

The Testing Sets. To be included in the testing set, a worker must have attempted at least 10 HITs (accepted or rejected) and have labeled at least 100 UI elements, so that a reasonably representative distribution of classes could be composed. Of all the recorded workers, only 20 (6.71%) have complied with this rule, but it was them who provided 272 (55.7%) of all accepted UIs and 17067 (53.9%) of all labeled elements.

In the subsequent sub-chapter which is dedicated to the investigation of our DGT method’s effectiveness in predicting performance in crowdworkers, we are exploring different sizes of the trusted set, ranging from 1 to 9. The trusted labelers are included to the trusted set of a particular size in the order defined by their quality index (see in Table 1): e.g. {VY, SV, KK} for size 3. The screenshots labeled by the included trusted labelers **are removed from the testing set**, while the remaining ones form the testing sub-set, for which the distribution and the workers’ performance are re-calculated. In other words, **there is never a redundancy**: in each setup, the sets of screenshots processed by the trusted labelers and the crowdworkers do not overlap. In Table 2, we show the descriptive statistics of the testing (sub-)sets (the number of labelers = 0 corresponds to the full testing set, which is included for reference only).

Table 2. The descriptive statistics of the testing sub-sets used in the DGT evaluation.

Trusted set			Testing (sub-)set			
# of labelers	# of UIs	% of UIs	# of workers	accepted HITs	rejected HITs	Precision (SD)
0	-	-	20	272	205	0.566 (0.453)
1	43	8.79%	19	253	175	0.595 (0.439)
2	87	17.72%	18	223	159	0.584 (0.454)
3	131	26.68%	17	201	127	0.621 (0.438)
4	175	35.64%	14	167	108	0.611 (0.436)
5	219	44.60%	12	145	72	0.709 (0.355)
6	262	53.25%	9	92	61	0.612 (0.408)
7	306	62.20%	5	71	8	0.900 (0.120)
8	360	74.53%	4	42	7	0.855 (0.145)
9	403	81.91%	2	22	0	1.000 (0.000)

3.2 The DGT Method Evaluation

Our DGT method relies on the assumption that the distribution of classes in UI labeling tasks is indicative of the overall performance, operationalized as Precision$_{AMT}$ in our study. In Figure 1 we show the distributions of the classes for the 20 workers of the tested set (AMT$_i$) and the overlaid distribution of the best trusted labeler (VY), which all resemble a power law distributions. For each of the workers, the classes were sorted by frequency, so the ranks on the horizontal axis may correspond to different classes. Then they were normalized by dividing each value by the mean frequency for the worker. For the trusted labeler, the same procedure was performed, and also every two frequencies were averaged to one rank and value in the diagram (as trusted labelers had 20 pre-defined classes instead of 10 classes for the workers).

The actual DGT method is straightforward now. For each tested worker, we take the distribution of the classes and apply the two-sample Kolmogorov-Smirnov test that compares it to the distribution of each labeler in the trusted set. The p-values produced by the tests are averaged for each worker and are used to predict precisions in the testing sub-set. So, we applied the method for all the sub-sets previously specified in Table 2 (even though with low number of workers in the testing sub-set the model makes little sense) and present the outcomes in Table 3. The best $R^2 = 0.855$ corresponding to the trusted set size = 2 (trusted labelers VY and SV) is highlighted. To evaluate our DGT model (with the best trusted set size = 2), we compared it to the baseline and considered some other alternative factors (see in Table 4). Particularly, we calculate R^2s in regressions built for Precision$_{AMT}$ for the following variables:

- ToT$_{AMT}$ – the baseline, often used in crowdsourcing quality control;
- attempted HITs (accepted + rejected) – the factor that nominally reflects the involvement of a worker in our study;

- EUI_{AMT} – the factor that arguably best reflects the actual work effort by a worker in our UI labeling HITs;
- GOF_{PL} – the goodness-of-fit measure of a worker's classes distribution to the power law distribution (obtained using the third-party `plpva.r` library that implements the test described in [20]).

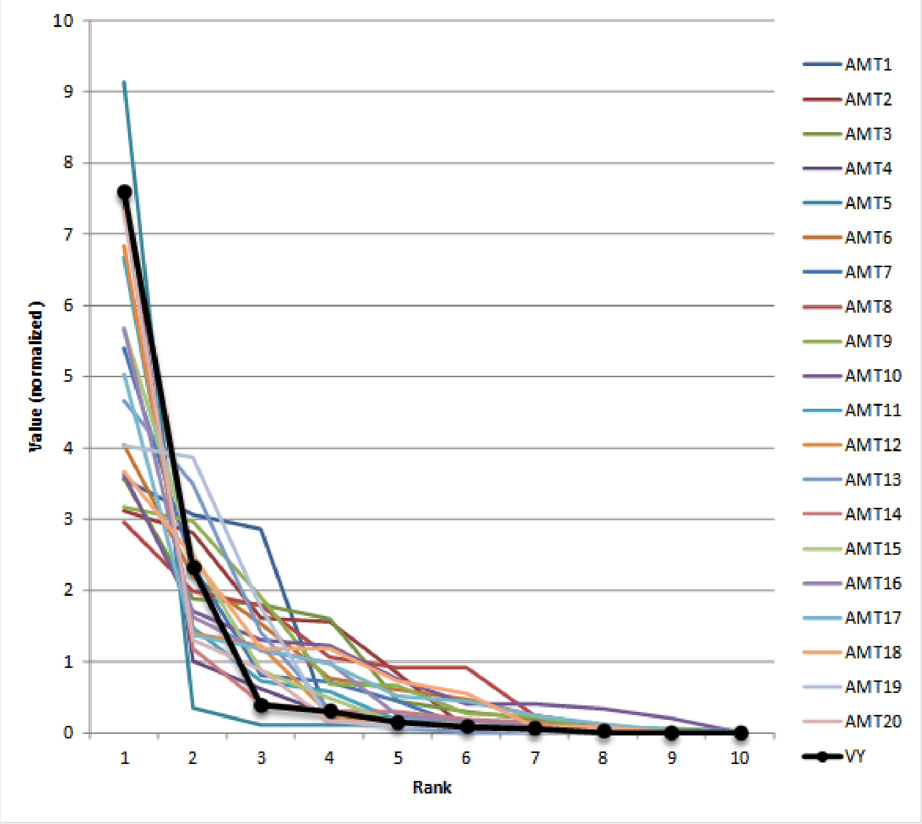

Fig. 1. The distributions of the labeling classes' frequencies for the crowdworkers (AMT_i) and the best trusted labeler (VY).

4 Discussion

The results of the DGT method evaluation imply that it might be applicable for the crowdsourcing data quality control in the UI labeling tasks that we considered. The R^2s produced by the method were of 0.8 and higher for the reasonably practical ratios between the trusted and the testing sets' sizes, 17-27% (Table 3). The ability of the method to predict performance in crowdworkers was considerably higher than that of the Time-on-Task factor (R^2 = 0.401) that is traditionally used for this purpose and that we considered as a baseline.

Table 3. The crowdworkers' performance prediction models resulting from the DGT method.

Size of trusted set	Size of testing sub-set	The model	
(# of trusted labelers)	(# of workers)	R^2	F, p
8.79% (1)	91.21% (19)	0.658	$F_{1,17} = 32.7$, $p < 0.001$
17.72% (2)	82.28% (18)	**0.855**	$F_{1,16} = 94.5$, $p < 0.001$
26.68% (3)	73.32% (17)	0.789	$F_{1,15} = 56.0$, $p < 0.001$
35.64% (4)	64.36% (14)	0.716	$F_{1,12} = 30.3$, $p < 0.001$
44.60% (5)	55.40% (12)	0.539	$F_{1,10} = 11.7$, $p = 0.007$
53.25% (6)	46.75% (9)	0.789	$F_{1,7} = 26.1$, $p = 0.001$
62.20% (7)	37.80% (5)	0.107	$F_{1,3} = 0.4$, $p = 0.591$
74.53% (8)	25.47% (4)	0.501	$F_{1,2} = 2.0$, $p = 0.292$
81.91% (9)	18.09% (2)	-	-

Meanwhile, an alternative factor EUIAMT that we also considered provided somehow superior R^2 compared to the DGT model's R^2s in some of the testing setups (Table 4). This is understandable, since EUI_{AMT} in our HIT was the best reflection of the work effort contributed to the task. We would however argue that the number of elements per UI is easily prone to malicious manipulations, similarly to the once indicative Time-on-Task. The latter in our study was inflated even by the workers who did not label a single UI element thus was not performing an actual task. Similarly, increased EUI_{AMT} could be futilely exaggerated with relatively little effort, e.g. through random specification of labels, possibly even with browser automation scripts. Also, in the trusted set that corresponds to higher-quality labeling data, the effect of EUI_T on either PrecisionT or Q was not significant, which questions the true impact of this factor.

The results presented in the table suggest that $R^2 = 0.875$ ($F_{1,16} = 112.0$, $p < 0.001$) for EUI_{AMT} was marginally higher than $R^2 = 0.855$ ($F_{1,16} = 94.5$, $p < 0.001$) for our DGT model, although somehow lower than $R^2 = 0.895$ ($F_{1,16} = 135.8$, $p < 0.001$) for one of the trusted labelers (SV) in the model. The GOF_{PL} factor was considerably less compelling ($R^2 = 0.480$, $F_{1,16} = 14.8$, $p = 0.001$), but still superior to the baseline ToT_{AMT} ($R^2 = 0.401$, $F_{1,16} = 10.7$, $p = 0.005$).

Another issue worth discussing is whether the effectiveness of the DGT method is due to the KS test considering mostly the locations of the distributions in our context. Indeed, the mean $EUI_T = 86.3$ was a great deal higher than the mean $EUI_{AMT} = 28.2$, and the mean number of UI elements in accepted HITs (58.3) would be closer to EUI_T. We however argue that the effect of precision in trusted labelers with respect to explaining the workers' $Precision_{AMT}$ was more prominent than the effect of completeness as expressed by SC_i. Indeed, of the considered two trusted labelers (see in Table 4), SV had higher $R^2 = 0.895$ than VY's $R^2 = 0.658$, notably lower $SC_{SV} = 80.4$ in comparison to $SC_{VY} = 95.5$, but higher $Precision_{SV} = 0.974$ vs. $Precision_{VY} = 0.928$. It might suggest that the effectiveness of the DGT method was mostly due to the distributions' shapes, though surely this statistically unrepresentative example calls for further investigation.

Table 4. The detailed results for the predictive model for PrecisionAMT (trusted set size = 2).

Precision_{AMT}	p-values from the KS test:			Alternative factors			
	with VY	with SV	Avg.	HITs	ToT_{AMT}	EUI_{AMT}	GOF_{PL}
0.974	0.856	0.837	0.847	39	191	56.31	0.492
0	0.091	0.158	0.124	34	50	4.62	0.617
1	0.276	0.937	0.606	32	558	70.13	0.706
0.813	0.686	0.987	0.837	32	325	37.16	0.640
0.731	0.974	0.704	0.839	26	102	24.00	0.589
0	0.002	0.002	0.002	25	63	5.24	0.115
0	0.066	0.012	0.039	23	126	4.39	0.354
0	0.458	0.158	0.308	19	77	9.21	0.562
0	0.019	0.023	0.021	19	94	6.11	0.406
1	0.482	0.517	0.499	18	619	57.72	0.640
1	0.686	0.704	0.695	18	232	39.06	0.600
1	0.608	0.875	0.741	16	1370	60.81	0.592
1	0.987	0.837	0.912	16	568	65.19	0.529
1	0.913	0.751	0.832	14	427	71.43	0.627
1	0.738	0.837	0.788	14	1326	68.43	0.639
0	0.259	0.032	0.146	14	57	7.21	0.296
1	0.913	0.751	0.832	12	837	76.83	0.659
0	0.003	0.010	0.006	11	355	9.27	0.431
R^2 for Precision_{AMT}:	**0.658**	**0.895**	**0.855**	**< 0.01**	**0.401**	**0.875**	**0.480**

The DGT method has certain inherent limitations. Arguably the strongest one is that a worker needs to produce enough results to compose a representative distribution of the classes – in our study, at least 100 UI elements labeled in 10 UIs. Indeed, the excluded workers contributed 216 (44.3%) of accepted HITs, which could not be covered by the method and would probably need to undergo different quality control procedures. However, one should consider that our experiment was artificially set up with a limited number of screenshots, whereas in real circumstances HIT design would be different. Moreover, the UI labeling task has an entry threshold – the workers need to comprehend the classes, read instructions, etc., so the learning effect is a positive thing and fewer workers each performing more HITs should be preferred to the contrary situation. Another limitation is that the trusted set might be bound to UIs belonging to a particular domain, and the transferability of the trusted distributions to other domains (e.g. from websites of universities to museums) is so far unexplored.

Finally, among the threats to validity we need to note that the assumptions for the two-sample KS test were not totally satisfied in our study. The variables (classes' distributions) were not continuous and the number of their values was rather modest (although ranging in a large interval). E.g. in [18] it is noted that for small sample sizes the nominal significance of 0.1 corresponds to the actual significance of 0.0835. However, the effects that we found in our study were rather strong and we probably can assume the findings are statistically valid. Also, we did try the `ks.boot` function in R that is considered an alternative to `ks.test` (adding simulation), but it did not produce any different p-values for our data.

5 Conclusions

The main contributions of our work can be summarized as follows:

- we proposed and evaluated the Distributional Ground Truth method for data quality control, which implies zero redundancy, thus having the potential to obviate the ancillary work effort and expenses;
- we demonstrated that shapes of classes' distributions (labels' frequencies) are reflective of the overall crowdworkers' performance in UI labeling tasks;
- we demonstrated advantage of comparing to a distribution obtained from a trusted dataset vs. a power law distribution, as previously done in [14].

At the current stage of research we must note that the boundaries for the DGT method applicability were not explored, neither has it been employed in real projects to assess the economic advantage over e.g. the widely used GT and MC methods due to the non-redundancy. In our context, one would probably have a GT of size 1, i.e. one completely and correctly labeled UI screenshot, and make sure that every worker has to label this, in a sort of an entrance-test to other HITs. Provided that in our AMT experimental session a worker on averaged labeled 4.17 screenshots, this would correspond to the quality control process leading to wasting 24% of the outcome. In case of MC implying that at least 3 workers label a screenshot, the share of the largely unused results would be even greater, at 67%.

In comparison to existing non-redundancy methods based on crowdworkers' behavior (foremost, on Time-on-Task), DGT provides an important advantage in "break-proofness", since it is based on regularities in true data. Thus, we see no uncomplicated way to imitate a trustworthy distribution of the classes for malicious workers even if they are aware of the employed data quality control method.

A practical issue is the desired number and quality of trusted labelers. Our assumption that trusted labelers with greater quality index would have "better" distributions was not confirmed in practice, as demonstrated by VY's and SV's R^2s in Table 4. So, we plan to explore efficient approaches for composing trusted sets for the DGT method in our further research work. Currently, we would just recommend having reasonable diversity of trusted labelers and assume that the averaged p-values negate the effect of individual discrepancies.

Our further research prospects include exploration of the method's applicability: whether it could be feasible in other crowdsourcing tasks, what are the efficient approaches for composing the trusted set, etc. However, even at the current stage of development we hope that our results can contribute to more efficient non-redundant crowd data quality control and thus to better utilization of human mind power in HCI-related ML tasks.

Acknowledgment. I would like to acknowledge Dr. Sebastian Heil, who contributed significantly to this research work, but whose name could not be included among the authors due to issues of international politics.

References

1. Chui, M. et al.: Notes from the AI frontier: Insights from hundreds of use cases. McKinsey Global Institute, 2 (2018).
2. Mao, K., Capra, L., Harman, M., & Jia, Y.: A survey of the use of crowdsourcing in software engineering. Journal of Systems and Software, 126, 57-84 (2017).
3. Bakaev, M., Heil, S., Gaedke, M.: A Reasonable Effectiveness of Features in Modeling Visual Perception of User Interfaces. Big Data and Cognitive Computing, 7(1), 30 (2023).
4. Naderi, B.: Motivation of workers on microtask crowdsourcing platforms (p. 125). Cham: Springer International Publishing (2018).
5. Alabduljabbar, R., Al-Dossari, H.: A dynamic selection approach for quality control mechanisms in crowdsourcing. IEEE Access, 7, 38644-38656 (2019).
6. Daniel, F. et al.: Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. ACM Computing Surveys (CSUR), 51(1), 1-40 (2018).
7. Heil, S., Bakaev, M., Gaedke, M.: Assessing completeness in training data for image-based analysis of web user interfaces. In CEUR Workshop Proceedings, 2500 (2019).
8. Schmidt, F. A.: Crowdsourced production of AI training data: How human workers teach self-driving cars how to see (No. 155). Working Paper Forschungsförderung (2019).
9. Baba, Y., Kashima, H.: Statistical quality estimation for general crowdsourcing tasks. In Proc of 19th ACM SIGKDD int conf on Knowledge discovery and data mining, pp. 554-562 (2013).
10. Le, J., Edmonds, A., Hester, V., Biewald, L.: Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In SIGIR 2010 workshop on crowdsourcing for search evaluation, 2126, pp. 22-32, (2010).
11. Fang, Y.L. et al.: Improving the quality of crowdsourced image labeling via label similarity. Journal of Computer Science and Technology, 32, 877-889 (2017).
12. ISO 13528:2015 Statistical methods for use in proficiency testing by interlaboratory comparison. International Organization for Standardization, Geneva.
13. Komarov, S., Reinecke, K., Gajos, K. Z.: Crowdsourcing performance evaluations of user interfaces. In Proc of the SIGCHI conf on human factors in comp systems, pp. 207-216 (2013).
14. Heil, S., Bakaev, M., Gaedke, M.: Web User Interface as a Message: Power Law for Fraud Detection in Crowdsourced Labeling. In Int Conf on Web Engineering, pp. 88-96 (2021).
15. Boychuk, E., Bakaev, M.: Entropy and compression based analysis of web user interfaces. In Web Engineering: 19th International Conference, ICWE 2019, Daejeon, South Korea, June 11–14, 2019, Proceedings 19 pp. 253-261 (2019).
16. Weidema, E.R. et al.: Toward microtask crowdsourcing software design work. In Proc of the 3rd Int Workshop on Crowdsourcing in Software Engineering, pp. 41-44 (2016).
17. Nebeling, M., Speicher, M., Norrie, M.C.: CrowdAdapt: enabling crowdsourced web page adaptation for individual viewing conditions and preferences. In Proc of the 5th ACM SIGCHI symposium on Engineering interactive comp systems, pp. 23-32 (2013).
18. Lemeshko, B.Y., Lemeshko, S.B.: On the convergence of distributed statistics and the power of Smirnov and Lehmann-Rosenblatt homogeneity criteria. Measurement Techn., 12, 9-14 (2005).
19. Gonzalez, T., Sahni, S., Franta, W. R. An efficient algorithm for the Kolmogorov-Smirnov and Lilliefors tests. ACM Transactions on Mathematical Software (TOMS), 3(1), 60-64 (1977).

20. Clauset, A., Shalizi, C. R., Newman, M. E.: Power-law distributions in empirical data. SIAM review, 51(4), 661-703 (2009).
21. Gordon, A.Y., & Klebanov, L.B.: On a paradoxical property of the Kolmogorov–Smirnov two-sample test. In Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis, 7, pp. 70-75). Institute of Mathematical Statistics (2010).