Estimating crowd-worker's reliability with interval-valued labels to improve the quality of crowdsourced work

1st Makenzie Spurling
Dept. of Comp. Sci. & Eng.
University of Central Arkansas
Conway AR, USA
mspurling1@cub.uca.edu

2nd Chenyi Hu
Dept. of Comp. Sci. & Eng.
University of Central Arkansas
Conway AR, USA
chu@uca.edu

3th Huixin Zhan
Dept. of Computer Science
Texas Tech University
Lubbock TX, USA
huixin.zhan@ttu.edu

4rd Victor S. Sheng
Dept. of Computer Science
Texas Tech University
Lubbock TX, USA
victor.sheng@ttu.edu

***Abstract*—With inputs from human crowds, usually through the Internet, crowdsourcing has become a promising methodology in AI and machine learning for applications that require human knowledge. Researchers have recently proposed interval-valued labels (IVLs), instead of commonly used binary-valued ones, to manage uncertainty in crowdsourcing [19]. However, that work has not yet taken the crowd worker's reliability into consideration. Crowd workers usually come with various social and economic backgrounds, and have different levels of reliability. To further improve the overall quality of crowdsourcing with IVLs, this work presents practical methods that quantitatively estimate worker's reliability in terms of his/her correctness, confidence, stability, and predictability from his/her IVLs. With worker's reliability, this paper proposes two learning schemes: weighted interval majority voting (WIMV) and weighted preferred matching probability (WPMP). Computational experiments on sample datasets demonstrate that both WIMV and WPMP can significantly improve learning results in terms of higher precision, accuracy, and F_1-score than other methods.**

***Index Terms*—crowdsourcing, interval-valued label, worker's reliability, correctness, confidence, stability, predictability**

I. Introduction

In this section, we briefly introduce the typical crowdsourcing problem that we study with a short literature review.

A. Crowdsourcing and worker's reliability

Crowdsourcing is the practice of obtaining inputs (labels) from a large number of people (crowd workers) typically via the Internet. Meaningful human knowledge from these labels can be further applied in machine learning and AI. Nowadays, large volumes of data manually labeled via human crowds have been collected such as Amazon Mechanical Turk[1], CrowdFlower[2], and others.

In this work, we assume the typical binary classification model in crowdsourcing. That is to determine if an observation $x \in X$ is an instance of a given class $y \in Y$. To simplify the discussion, researchers often assume $|Y| = 1$ without loss of generality. This is because when $|Y| = m > 1$, one may check if x_i is an instance of each of the m classes repeatedly. With the assumption of $|Y| = 1$, the problem becomes a decision problem: is $x_i \in X$ an instance of a given class y? The correct answer, either yes (1) or no (0), is the ground truth. The basic idea of crowdsourcing is to acquire a list of answers (labels), denoted as L_i, from a crowd on the same x_i. Applying a learning strategy to aggregate L_i, one makes an inference with the objective of matching the ground truth. Obviously, the overall quality of collected labels plays a critical role in addition to the learning strategy.

Due to the open nature, crowd workers usually come with diverse social-economical backgrounds. Workers with a higher level of domain knowledge usually make better quality labels than those with less. However, an expert with adversarial purposes can cause more harm than good. Even without any adversarial intention, experts may often disagree [26]. In addition, cultural and demographic differences often lead to biased labels. Even for the same worker, he may not perform exactly the same all time because of variations of his emotional and stress level.

In short, crowd workers are usually not equally reliable. To improve the quality of crowdsourced work, we should take worker's reliability into consideration.

B. Related previous work and motivation of this study

Researchers have studied worker's reliability previously. In [1], Bi *et al* explicitly studied sources for a noisy label including worker's dedication, expertise, default labeling judgment, and sample difficulty. In [23] and [27], the authors discussed effective and efficient ways to select a subsets of workers to maximize the accuracy under a budget constraint. Qiu *et al* proposed ways to select worker through behavior prediction [33]. Wang *et al* reported practical strategies for adversarial detection in [35]. In [30], the correctness of a worker j is described as a probability p_j obtained from historic data. Applying the probability with EM algorithm [5], one may improve the quality of statistical inferences. Tao *et al*

This work is partially supported by the US National Science Foundation through the grant award NSF/OIA-1946391.

[1]https://www.mturk.com/

[2]http://crowdflowersites.com/

considered worker's reliability in MV-Freq and MV-Beta [32], and reported quality improvements in [34].

Previous studies in crowdsourcing mostly use binary-valued labels. However, a worker may often have ambiquity when having to select 0 or 1 definitely in practice. Uncertainties like this are inevitable when gathering crowdsourced data. This is the reason to use interval-valued labels over binary ones. Forcing the worker to use binary-valued labels loses the uncertainity a worker has in their inputs. The more information gained from the workers, the better the algorithms and models predictions. To include this information, Hu *et al* introduced interval-valued labels (IVL) in [19]. Intervals have their own specific properties and operations. With them, people have made significant progresses in solving otherwise hard problems [3], [4], [6], [7], [9]–[18], [20]–[22], [24], [25], [28], and more. Likewise, computational results in [19] evidence quality improvements with IVLs than without. However, that work implicitly treats all IVLs equally without considering worker's reliability. Noticing IVLs from a worker j contain information on j's reliability, we study worker's reliability with IVLs in this work to further improve the quality of crowdsourced work.

The rest of this paper is organized as follows. In section 2, we briefly introduce related background knowledge. In section 3, we quantitatively study ways to estimate worker's reliability from his IVLs in terms of his overall correctness, confidence, stability, and predictability. In section 4, we present strategies that apply worker's reliability on worker selection, and inference making. We report results of computational experiments in section 5; and summarize the work in section 6.

II. Background knowledge and notations

In this section, we introduce notations and properties of IVLs after a very brief review on binary-valued labels on gold questions.

A. Binary-valued labels on gold questions

To computationally estimate worker's correctness, a common approach in the literature is to employ a set of *gold questions*. For each gold question, its ground truth is known but opaque to workers. Let $G = [g_1, g_2, \ldots, g_k]$ be a list of gold questions. Then, the ground truth of G is a binary string $o(G) \in \{0,1\}^k$. Let L_G^j be the list of binary-valued labels on G by a worker j. For each $g_i \in G$, $l_i^j \in L_G^j$ may or may not match the ground truth $o(g_i)$. Comparing the binary string L_G^j against $o(G)$, one may predict j's correctness. For example, if the count of total matching is c_j, then the ratio $p_j = c_j/k$ is an empirical probability of j's overall correctness. Assuming j labels each $g_i \in G$ independently with exactly the same probability of success p_j, then L_G^j records the result of Bernoulli trials. Solving Eq. (1) numerically, one may find the probability p_j:

$$\frac{c_j}{k} = \left(\begin{array}{c} k \\ c_j \end{array} \right) p_j^k (1-p_j)^{k-c_j}. \tag{1}$$

People have applied the p_j to quantify j's reliability. However, the assumptions of independence and the same probability of success may not always hold in real applications. A binary-valued label may also cause information loss, especially when j has ambiguity in selecting either 0 or 1 definitely. The lost information is closely related to j's reliability, and hard to recover through post processing. In contrast to binary-valued labels, IVL allows j to specify his uncertainty. With such additional information in j's IVLs, we can study j's reliability in terms of his correctness, confidence, stability, and predictability in this section.

B. Properties of IVLs and notations used in this work

Prior to our discussion, let us clarify some notation rules first. In the literature of interval computing, people often denote an interval object with a boldface letter to distinguish it from a real valued one (not boldface). The greatest lower and least upper bounds of an interval object are specified with an underline and an over-line of the same letter without boldface, respectively. Hence, the IVL for an instance i made by j is denoted as $\boldsymbol{l}_{ij} = [\underline{l}_{ij}, \overline{l}_{ij}] \subseteq [0,1]$. The minimum and maximum beliefs of j on i being an instance of the given class are $\underline{l}_{ij}$ and $\overline{l}_{ij}$, respectively. The midpoint or centroid of $\boldsymbol{l}_{ij}$ is point-valued. We write it as

$$\mathrm{mid}(l_{ij}) = \frac{\underline{l}_{ij} + \overline{l}_{ij}}{2} \tag{2}$$

without boldface $\boldsymbol{l}_{ij}$ because $\mathrm{mid}(l_{ij})$ is a real. Because $\boldsymbol{l}_{ij} \subseteq [0,1]$, we have $0 \leq \mathrm{mid}(\boldsymbol{l}_{ij}) \leq 1$. When $\mathrm{mid}(l_{ij}) > 0.5$, j leans toward to accepting x_i in the class. We call it a positive IVL. If $\mathrm{mid}(l_{ij}) < 0.5$, then j leans toward rejecting x_i from the class. We call it a negative IVL. Otherwise, it is neither positive nor negative, and implies a tie. The radius of $\boldsymbol{l}_{ij}$,

$$\mathrm{rad}(l_{ij}) = \frac{\overline{l}_{ij} - \underline{l}_{ij}}{2}, \tag{3}$$

is point-valued too. So, the l_{ij} is not in boldface. The radius of $\boldsymbol{l}_{ij}$ specifies the range of variations from the centroid.

Let $\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_n]$ be a list of IVLs. Both its lower and upper bounds $\underline{L}$ and $\overline{L}$ are real vectors without boldface. So are the element-wise midpoint and radius vectors. We denote them as $\mathrm{mid}(L)$ and $\mathrm{rad}(L)$, respectively, without boldface. The mean of $\boldsymbol{L}$ is an interval as the following

$$\mu(\boldsymbol{L}) = \frac{\sum_{i=1}^n \boldsymbol{l}_i}{n} = \left[\frac{1}{n}\sum_{i=1}^n \underline{l}_i, \frac{1}{n}\sum_{i=1}^n \overline{l}_i \right] = [\mu(\underline{L}), \mu(\overline{L})]. \tag{4}$$

The variance of $\boldsymbol{L}$ derived in [17] is a real value denoted as

$$Var(L) = Var(\mathrm{mid}(L)) + Var(\mathrm{rad}(L)) + \frac{2}{n}\sum_{i=1}^n |\Delta m_i \Delta r_i| \tag{5}$$

where $\Delta m_i = \mathrm{mid}(l_i) - \mu(\mathrm{mid}(L))$ and $\Delta r_i = \mathrm{rad}(l_i) - \mu(\mathrm{rad}(L))$. Hence, the standard deviation of $\boldsymbol{L}$ is

$$\sigma(L) = \sqrt{Var(L)}. \tag{6}$$

We say that a function $f(t)$ is a *pdf* if and only if

$$\left\{ \begin{array}{l} f(t) \geq 0 \ \forall t \in (-\infty, \infty), \text{ and} \\ \int_{-\infty}^{\infty} f(t)dt = 1. \end{array} \right. \tag{7}$$

Eq. (8) provides a *pdf* (probability density function) for $\boldsymbol{L}$:

$$f(t) = \frac{\sum_{i=1}^{n} pdf_i(t)}{n}, \tag{8}$$

where pdf_i is a *pdf* of a random variable $l_i \in \boldsymbol{l}_i$.

In the rest of this paper, we apply the above statistic and probabilistic properties of IVLs to study worker's reliability. We use $\boldsymbol{L}_i$ to denote the list of IVLs on the same instance i by some workers $j \in J$. We use $\boldsymbol{L}^j$ to denote the list of IVLs made by the same worker j on different observations. When needed, we use $\boldsymbol{L}^j_G$ and $\boldsymbol{L}^j_X$ to distinguish j's IVLs on G (a set of gold questions) and X (regular questions), respectively.

III. Estimating worker's reliability from his IVLs

In this section, we quantitatively specify worker's reliability from his IVLs in terms of his correctness, confidence, stability, and predictability.

A. Estimating worker's correctness from his IVLs on a set of gold questions

To study j's correctness, we collect his IVLs on G. The IVL from j on a $g \in G$ is denoted as $\boldsymbol{l}_{gj} = [\underline{l}_{gj}, \overline{l}_{gj}]$. With the known ground truth $o(g)$, we have the *center-correctness* of $\boldsymbol{l}_{gj}$ represented in its centroid as:

$$\text{center\_correctness}(l_{gj}) = \begin{cases} 1 - \text{mid}(l_{gj}) & \text{if } o(g) = 0, \\ \text{mid}(l_{gj}) & \text{if } o(g) = 1. \end{cases} \tag{9}$$

The center-correctness of $\boldsymbol{l}_{gj}$ relies on both $\text{mid}(l_{gj})$ and $o(g)$. To simplify our discussion, we assume $o(g) = 1$ for all $g \in G$ without loss of generality. This is because the value of $o(g)$ is known. In the case $o(g) = 0$, we can replace $\boldsymbol{l}_{gj}$ with its difference from 1, i.e. $1 - \boldsymbol{l}_{gj} = [1 - \overline{l}_{gj}, 1 - \underline{l}_{gj}]$ without changing its center-correctness.

For example, let $o(g) = 0$ and $\boldsymbol{l}_{gj} = [0.2, 0.4]$, then the center-correctness is $1 - 0.3 = 0.7$. Converting $o(g)$ to 1 and replacing $\boldsymbol{l}_{gj}$ with $[1 - 0.4, 1 - 0.2] = [0.6, 0.8]$, Eq. (9) gives exactly the same center-correctness of 0.7.

Hereafter, unless specified otherwise, we assume $o(g) = 1$ for all gold questions upon the replacement of $\boldsymbol{l}_{gj}$ with $1 - \boldsymbol{l}_{gj}$ whenever $o(g) = 0$ originally. By doing so, the center-correctness of an IVL on a gold question g by j is $\text{mid}(l_{gj})$. Similar to center-correctness, we call the values of $\underline{l}_{gj}$ and $\overline{l}_{gj}$ the *min-* and *max-correctness* of the label $\boldsymbol{l}_{gj}$.

Let $\boldsymbol{L}^j_G = [\boldsymbol{l}_{g_1 j}, \boldsymbol{l}_{g_2 j}, \ldots, \boldsymbol{l}_{g_k j}]$ be the list of k IVLs from j on G. Then, the mean of $\boldsymbol{L}^j_G$ is $\mu(\boldsymbol{L}^j_G) = [\mu(\underline{L}^j_G), \mu(\overline{L}^j_G)]$. It provides estimations of the overall correctness of j in terms of his average min-, max-, and center-correctness $\mu(\underline{L}^j_G), \mu(\overline{L}^j_G)$, and $\text{mid}(\mu(L^j_G))$, respectively. Contrast to the probability p_j derived from Eq. (1), the average correctness of j represented in $\mu(\boldsymbol{L}^j_G)$ does not require the strict assumptions of Bernoulli trial. Furthermore, the average min-, max-, and center-correctness of j provide us the means of the worst, best, and average correctness of j. Moreover, the standard deviations of $\underline{L}^j_G, \overline{L}^j_G$, and $\text{mid}(L^j_G)$ provide information on the stability of j's min-, max-, and center-correctness, respectively. This means that $\boldsymbol{L}^j_G$ contains more information about j's correctness than the p_j in Eq. (1).

B. Estimating the confidence of a worker j from his IVLs

An IVLs $\boldsymbol{l} = [\underline{l}, \overline{l}]$ contains information of labeler's confidence. The centriod of $\boldsymbol{l}$, $\text{mid}(l)$, represents the degree of the worker's belief toward 0 or 1. When $\text{mid}(l) = 0.5$, the worker has absolutely no confidence to pick either 0 or 1. The distance between $\text{mid}(l)$ and 0.5, i.e. $|\text{mid}(l) - 0.5|$, reflects the labeler's confidence on his belief. The radius of $\boldsymbol{l}$, $\text{rad}(l) = \frac{\overline{l} - \underline{l}}{2}$, specifies the maximum possible variation from the centroid. When $\text{rad}(l) = 0$, l is point-valued; and the worker is confident on the value of l. Otherwise, the label $\boldsymbol{l}$ contains uncertainty over a range. Because the maximum possible value of $\text{rad}(l)$ is 0.5, the difference between 0.5 and $\text{rad}(l)$, i.e. $0.5 - \text{rad}(l)$, measures labeler's confidence on the centroid. We say the confidence of $\boldsymbol{l}$ is

$$\text{conf}(l) = |\text{mid}(l) - 0.5| + 0.5 - \text{rad}(l). \tag{10}$$

Because both of $|\text{mid}(l) - 0.5|$ and $0.5 - \text{rad}(l)$ are between 0 and 0.5, the confidence of $\boldsymbol{l}$ is between 0 and 1. In contrast, the confidence of any binary-valued label is 100%. This is because, for a binary-valued label $l = 0$ (or 1), $|\text{mid}(l) - 0.5| = 0.5$ and $\text{rad}(l) = 0$. This means that the confidence of a binary-valued label is not distinguishable. In contrast, we are able to differentiate IVLs with their confidence values. For instance, the confidence values of $[0.8, 0.9]$ and $[0.5, 0.7]$ are $0.8 = 0.35 + 0.45$ and $0.5 = 0.1 + 0.4$, respectively.

As mentioned earlier, the mean of $\boldsymbol{L}^j_G$, $\mu(\boldsymbol{L}^j_G)$, is a sub-interval of [0, 1]. It not only provides j's average correctness, but also reflects j's overall confidence as $|\text{mid}(\mu(L^j_G)) - 0.5| + 0.5 - \text{rad}(\mu(L^j_G))$. We want to make two clarifications here. The first is that j's overall confidence is not the same as the mean of conf($\boldsymbol{l}_{g_i j}$). The other one is that unlike predicting worker's correctness, estimating worker's confidence with Eq. (10) does not require the ground truth but only the mean of IVLs. With this in mind, we can calculate j's overall level of confidence on X. Comparing j's confidences $\text{conf}(L^j_G)$ and $\text{conf}(L^j_X)$, we may statistically check if j performs consistently or not. If G well samples X, then $\text{conf}(L^j_G)$ and $\text{conf}(L^j_X)$ should be statistically consistent.

C. Estimating worker's stability and predictability from his IVLs

The standard deviation of a data set measures its overall stability. As mentioned earlier, j's min-, max-, and center-correctness as well as his confidence are point-valued. We can calculate their standard deviations as usual. Beyond that, Eqs. (5) and (6) enable us to calculate $\sigma(L^j)$ from $\boldsymbol{L}^j$ to estimate j's overall stability. Similar to j's confidence, the standard deviation of $\boldsymbol{L}^j$ does not rely on the ground truth. So, we can calculate and compare $\sigma(L^j_G)$ and $\sigma(L^j_X)$ statistically.

To measure j's predictability, we apply the entropy of $\boldsymbol{L}^j$ according to information theory. Let s be a discrete random variable with possible outcomes $s_1, s_2, \ldots, s_n$, which occur

with probability $\mathrm{p}(s_1), \mathrm{p}(s_2), \ldots, \mathrm{p}(s_n)$. Then, Shannon's entropy [29] of s is:

$$H(s) = -\sum_{i=1}^{n} p(s_i) \log p(s_i). \tag{11}$$

Assume $|\boldsymbol{L}^j| = m$ and pdf^j are the pdf of a random variable l over an $\boldsymbol{l}^j \in \boldsymbol{L}^j$. The $2m$ endpoints in $\boldsymbol{L}^j$, though some of them may overlap, form a partition of the interval $[0, 1]$ into $2m + 1$ sub-intervals. With Eq. (8), we are able to calculate the probability of each of the $2m + 1$ sub-intervals with Algorithm 1 in [17]. We can then apply Eq. (11) to obtain $H(L^j)$, the entropy of $\boldsymbol{L}^j$ as a quantitative measure on j's predictability. According to the minimum entropy principle, a worker j is more predictable if the $H(L^j)$ is less than others'. The calculation of entropy does not depend on the ground truth of instances to be classified. For a specific $j \in J$, we can calculate and compare the values of $H(L_G^j)$ and $H(L_X^j)$ statistically as well.

IV. Applying Worker's Reliability to Worker Selection and Inference Making

With IVLs from j, we are able to quantitatively measure his reliability in terms of correctness, confidence, stability, and predictability. Now, we apply worker's reliability to worker selection and inference making.

A. Worker selection

The purpose of worker selection is to improve the quality of crowdsourced work. Previous studies in worker selection mostly apply worker's correctness estimated from binary-valued labels on gold questions [23], [33], etc. With $\boldsymbol{L}_G^j$, we have newly estimated j's correctness together with his confidence, stability, and predictability available for worker selection. For example, the confidence of $\mu(\boldsymbol{L}_G^j)$ may differentiate j from others.

To combine j's correctness and confidence together in worker selection, we first investigate the relation between j's correctness and confidence. The estimated j's center-correctness from $\mu(\boldsymbol{L}_G^j)$ is $\mathrm{mid}(\mu(L_G^j))$, which is between 0 and 1. The radius of $\mu(\boldsymbol{L}_G^j)$, $\mathrm{rad}(\mu(L_G^j))$, is between 0 and $\min\{\mathrm{mid}(\mu(L_G^j)), 1 - \mathrm{mid}(\mu(L_G^j))\}$ to ensure $\mu(\boldsymbol{L}_G^j) \subseteq [0, 1]$. According to Eq. (10), the range of confidence for any given $\mathrm{mid}(\mu(L_G^j))$ is $|\mathrm{mid}(\mu(L_G^j)) - 0.5| + 0.5 - \min\{\mathrm{mid}(\mu(L_G^j)), 1 - \mathrm{mid}(\mu(L_G^j))\}$.

Fig. 1 visually illustrates the range of confidence vs. a given center-correctness. It suggests that a worker with a high level of center-correctness has a high level of confidence too. For example, when j's overall center-correctness is above 90%, his level of confidence is at least 80%. However, the converse is not true. Given j's confidence level at 80% or above, the range of his center-correctness can be less than 20% or above 80%. While it is good to have a worker with correctness above 80%; it seems unacceptable if the correctness is less than 20%. However, because of the binary classification model, we may still utilize labels from j, even though his correctness is less than 20%. Replacing his $\boldsymbol{l}_{ij}$ with $1 - \boldsymbol{l}_{ij}$, we would expect

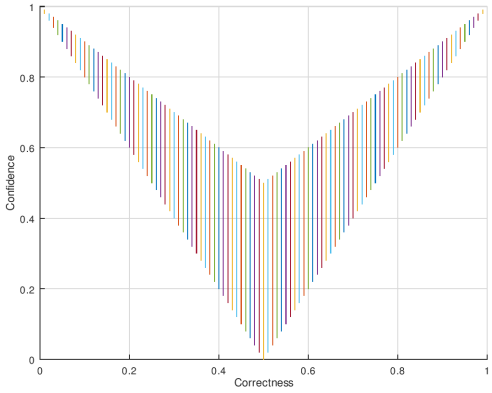

Fig. 1: Range of confidence vs. correctness

an average center-correctness above 80%. Another observation from Fig. 1 is that the range of center-correctness converges to 0.5 as confidence approaches 0. This implies that a label from a worker with a very low confidence is likely a tie. Summarizing the above discussion, we suggest the following heuristics for worker selection:

- A worker j is preferable if he has a high level of confidence above a threshold. For example, if the confidence threshold is 80%, then the average correctness can be above 80% too upon the difference from 1 replacement when $\mathrm{mid}(\mu(L_G^j)) \leq 0.2$.
- When $\mu(\boldsymbol{L}_G^j)$ has a mediocre confidence (say between 40-60%), then j's correctness may vary in a rather broad range. For instance, if the confidence level is 40%, then the center-correctness can be between 30-70%. If we need to select a worker with a mediocre confidence, the one with high correctness would be preferable.
- A worker j is not very helpful if his confidence level is very low. For instance, if j's confidence is below 20%, then the average of his IVLs is at most 10% away from a tie (50%).

The confidence of j providse us an additional criterion, other than correctness, for worker selection.

A challenging task in worker selection is to identify and exclude those who are very knowledgeable but have adversary purposes [2], [33]. A naive attacker with good knowledge may purposely classify all instances opposite to the ground truth. His answers on the gold questions result in a very low level of correctness. By replacing his labels with the difference from 1, we are still able to utilize his labels. In contrast to a naive attacker, a sophisticated attacker may be able to identify gold questions and label them correctly in order to be selected. After that, he launches attacks when answering regular questions. In practice, it has been suggested to monitor workers with very high correctness. A very sophisticated attacker may manipulate his correctness in answering gold questions to pass the threshold and avoid being identified. When a worker tries to lower his correctness with IVLs, his confidence level changes. Gold questions are not

required when calculating worker's confidence, stability, and predictability. This means that if G well samples X, then the overall confidence, stability, and entropy derived from $\boldsymbol{L}_G^j$ and $\boldsymbol{L}_X^j$ should be statistically consistent. This can be helpful in identifying possible attackers, which is an important subject in crowdsourcing. However, due to page limitation, we will discuss this in detail later in another paper.

B. Applying worker's reliability in inference making with improved matching probability

The objective of crowdsoucing is to derive an inference from collected labels that matches the ground truth. Let $\boldsymbol{L}_i = \{\boldsymbol{l}_{i1}, \boldsymbol{l}_{i2}, \ldots, \boldsymbol{l}_{im}\}$ be the IVLs collected from m workers on the same $x_i \in X$. Two strategies are suggested in [19] on making an inference from $\boldsymbol{L}_i$. One of them mimics the majority voting (MV) with binary-valued labels. It counts the numbers of positive and negative IVLs, denoted as c_i^+ and c_i^-, according to if $\mathrm{mid}(l_{ij})$ is greater or less than 0.5. By comparing c_i^+ and c_i^-, the inference is

$$y_i = \begin{cases} 1 & \text{if } c_i^+ > c_i^- \\ 0 & \text{if } c_i^+ < c_i^- \\ tie & \text{otherwise} \end{cases} \tag{12}$$

Using binary-valued labels, a reasonable worker would label x_i with 1 (or 0) if $\mathrm{mid}(l_{ij})$ is greater (or less) than 0.5. Hence, Eq. (12) leads to an inference the same as that of binary-valued labels. A straightforward modification of Eq. (12) is to not count each positive (or negative) IVL as one but with its centroid $\mathrm{mid}(l_{ij})$ instead. That is

$$W_i^+ = \sum_{\boldsymbol{l}_{ij} \in \boldsymbol{L}_i \wedge \mathrm{mid}(l_{ij}) > 0.5} \mathrm{mid}(l_{ij}), \text{ and} \tag{13}$$

$$W_i^- = \sum_{\boldsymbol{l}_{ij} \in \boldsymbol{L}_i \wedge \mathrm{mid}(l_{ij}) < 0.5} 1 - \mathrm{mid}(l_{ij}). \tag{14}$$

Then, we have IMV (interval MV) as

$$y_i = \begin{cases} 1 & \text{if } W_i^+ > W_i^-, \\ 0 & \text{if } W_i^+ < W_i^-, \\ tie & \text{otherwise.} \end{cases} \tag{15}$$

However, not all labels are created equal. Some labels should be weighted more than others [36]. Using j's reliability r_j as the weight of his label $\boldsymbol{l}_{ij}$, we have

$$W_i^+ = \sum_{\boldsymbol{l}_{ij} \in \boldsymbol{L}_i \wedge \mathrm{mid}(l_{ij}) > 0.5} r_j \times \mathrm{mid}(l_{ij}), \text{ and} \tag{16}$$

$$W_i^- = \sum_{\boldsymbol{l}_{ij} \in \boldsymbol{L}_i \wedge \mathrm{mid}(l_{ij}) < 0.5} r_j \times (1 - \mathrm{mid}(l_{ij})). \tag{17}$$

Using Eqs. (16) and (17) instead of (13) and (14) to make an inference with Eq. (15), we have WIMV (weighted IMV) to make an inference from $\boldsymbol{L}_i$. The strength of the inference with WIMV (or IMV) is

$$\hat{p} = \max\left\{ \frac{W_i^+}{W_i^+ + W_i^-}, \frac{W_i^-}{W_i^+ + W_i^-} \right\}. \tag{18}$$

In the discussion above, we ignore such $\boldsymbol{l}_{ij} \in \boldsymbol{L}_i$ whenever $\mathrm{mid}(l_{ij}) = 0.5$.

The other inference scheme in [19] applies a *pdf* of $\boldsymbol{L}_i$ defined as

$$f_i(t) = \frac{\sum_{j=1}^m pdf_{ij}(t)}{m}, \tag{19}$$

where pdf_{ij} is a *pdf* of $\boldsymbol{l}_{ij}$. Because any value $t > 0.5$ implies a preference of 1, the probability of the overall preference of 1 on x_i is

$$P^+(0.5) = \int_{0.5}^1 f_i(t)dt. \tag{20}$$

Eq. (21) results in an inference with a preferred matching probability (PMP) as the following

$$y_i = \begin{cases} 1 & \text{if } P^+(0.5) > 0.5, \\ 0 & \text{if } P^+(0.5) < 0.5, \text{ and} \\ tie & \text{otherwise.} \end{cases} \tag{21}$$

Computational experiments reported in [19] clearly demonstrate that inferences from the same $\boldsymbol{L}_i$ with PMP have much better overall quality than those with MV.

Notice that the arithmetic average in Eq. (19) treats all pdf_{ij} equally without considering worker's reliability. To include worker's reliability in calculating f_i, we multiply j's reliability r_j with pdf_{ij} as a weight, and have

$$f_i(t) = \frac{\sum_{j=1}^m r_j \times pdf_{ij}(t)}{\sum_{j=1}^m r_j}. \tag{22}$$

It is straightforward to verify that the $f_i(t)$ in Eq. (22) is a *pdf* of $\boldsymbol{L}_i$ too. Applying it in Eq. (20), we can evaluate P^+ then make an inference with Eq. (21). We call it WPMP (weighted PMP). The probability of matching is $\hat{p} = \max\{P^+, 1 - P^+\}$ for both PMP and WPMP.

We have one more question. Which value should we use as the weight r_j? The objective of learning is to make an inference that matches the ground truth. A worker with high confidence, stability, and predictability may miss the ground truth completely. Among the four reliability indicators, only the correctness is directly associated with ground truth. Hence, we should use j's correctness as the weight r_j. As demonstrated in our computational experiments below, using correctness as weight in WIMV and WPMP, we can significantly improve the overall quality.

V. Computational experiments

In this section, we report our computational experiments to examine if we achieve quality improvements with considerations of worker's reliability in crowdsourcing.

A. The design of our experiments

To test the concepts and methodologies discussed, we generate a pool of workers randomly with various levels of reliability. Fig. 2 illustrates the pool of workers in clusters according to their correctness and confidence. From the pool, we randomly select a group of at least ten workers whose levels of confidence are no less than a preset threshold. These selected workers are asked to provide random IVLs on a

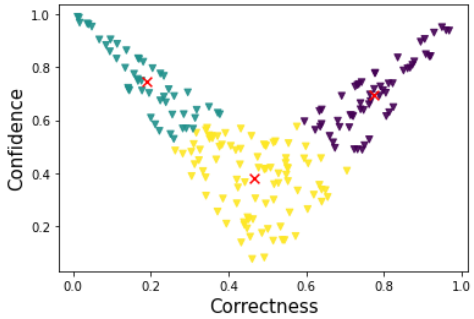

Fig. 2: A pool of workers with different reliability

Strategy	TP	FP	TN	FN
MV	24	48	37	39
IMV	34	56	44	46
PMP	25	53	39	48
WIMV	74	3	97	6
WPMP	75	1	99	5

TABLE I: Confusion matrices for strategies without correction

benchmark dataset in CEKA [38]. We then apply MV, IMV, PMP, WIMV, and WPMP to make inferences from these collected IVLs, and then check against the known ground truth. For the inferences obtained with each of the strategies, we record its confusion matrix [37] for the purpose of quality comparison. Table I records the results of a single run on the Income94 dataset with 180 items that used a worker confidence threshold of 60% and no label correction, i.e. without the $1 - \boldsymbol{l}_{ij}$ replacement of $\boldsymbol{l}_{ij}$ when j's correctness is very low. The table indicates that both WIMV and WPMP produce better results in terms of significantly reduced false positive (FP) and false negative (FN), while IMV and PMP miss a majority of ground truth. In addition, MV classified only 148 items with the remaining 32 being marked as a tie.

For comparison, Table II shows the results of a another test on the same dataset using the same confidence threshold 60% and the same workers but utilizing label correction via replacing $\boldsymbol{l}_{ij}$ with $1 - \boldsymbol{l}_{ij}$ when j's confidence is greater than 90%. In this table, all five strategies produce better results than those reported in Table I. This is due to utilizing label correction for high confidence, low correctness workers.

Tables I and II indicate that both WIMV and WPMP can produce much better results with random workers whose confidence level is at least 60%. It is impractical to verify the findings with various confidence thresholds through counting the numbers on confusion matrices. Instead, we apply the metrics of recall, precision, accuracy, and F_1-score to quantitatively measure the performance.

Strategy	TP	FP	TN	FN
MV	45	30	55	18
IMV	60	30	70	20
PMP	49	35	62	23
WIMV	79	0	100	1
WPMP	78	2	98	2

TABLE II: Confusion matrices for strategies with correction

B. Computational results

We implemente the design in Python 3 and illustrate the results graphically. Figs. 3-6 illustrate the changes of recall, precision, accuracy, and F_1-score for the Income94 dataset as the confidence threshold increases by 1% for each of the inference strategies. For each confidence threshold, ten workers are randomly selected from the pool in each run. The scores are then averaged over forty (no specific reasons) runs. Labels for workers with a correctness lower than 20% were replaced with $1 - \boldsymbol{l}_{ij}$.

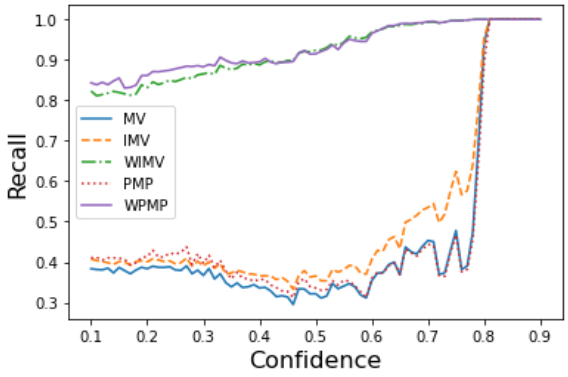

Fig. 3: Recall values vs. confidence threshold

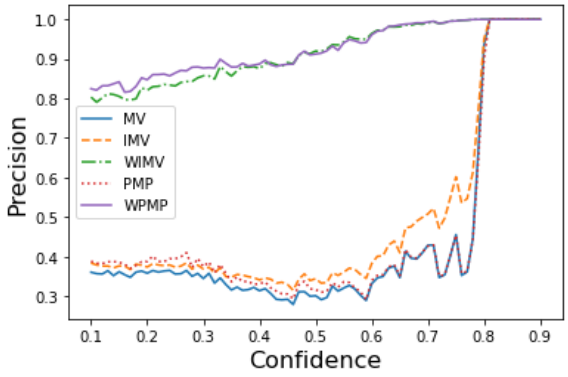

Fig. 4: Precision values vs. confidence threshold

Figs. 3-6 shows the average recall, precision, accuracy, and F_1-score values with an increasing confidence threshold on the dataset. All four figures show similar properties. When the confidence threshold is less than 60%, the qualitative measures on inferences made with MV, IMV,and PMP are all below 50%. When the confidence threshold increases beyond 60%, the measures show an increasing trend. Nevertheless, the results produced with WIMV and WPMP are above 0.8 even at a low confidence threshold. The significantly improved overall quality of WIMV and WPMP comes from the reliability weighted strategies proposed in this work.

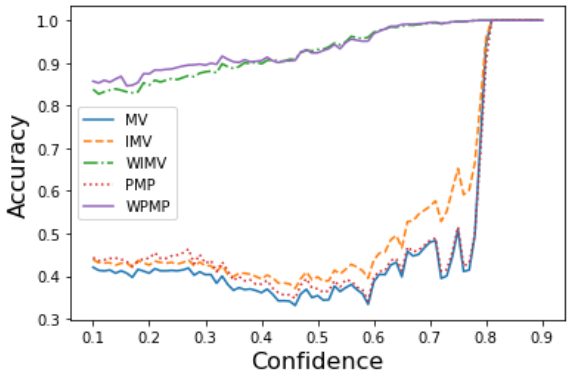

Fig. 5: Accuracy values vs. confidence threshold

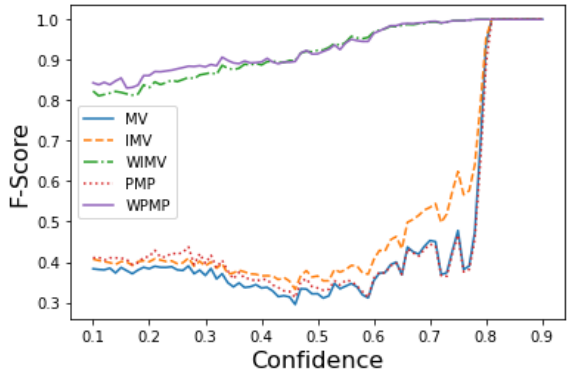

Fig. 6: F_1-score values vs. confidence threshold

To verify the findings, we ran the same experiments with the same workers on another benchmark dataset named car in CEKA. Figs. 7–10 report the qualitative measures. They show very similar properties to Income94, with the only difference being that precision and F-score start relatively lower than others when the confidence threshold is low. Overall, these

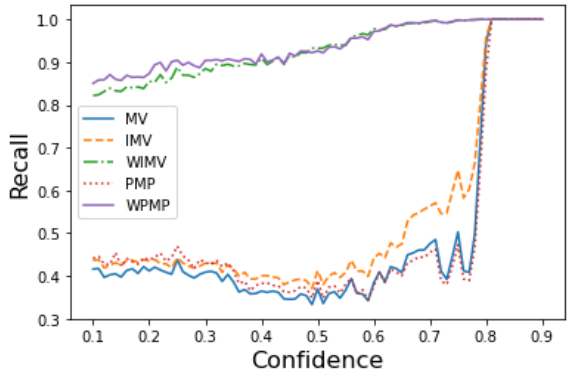

Fig. 7: Recall values vs. confidence threshold

figures demonstrate a significant improvement of the reliability weighted strategies, i.e. WIMV and WPMP, especially when the threshold of worker's confidence is relatively low. Using worker's confidence for worker selection is also shown to be

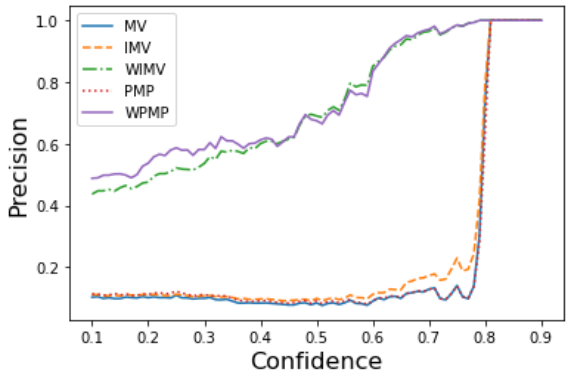

Fig. 8: Precision values vs. confidence threshold

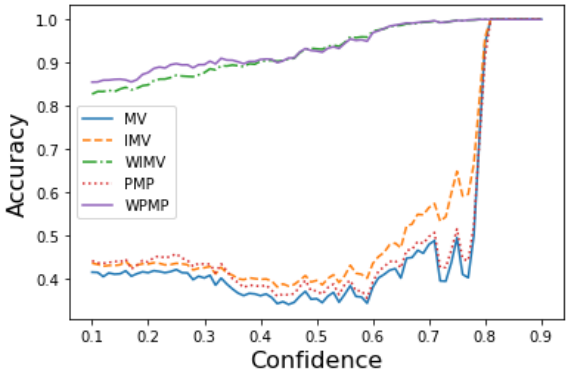

Fig. 9: Accuracy values vs. confidence threshold

an effective method to get an accurate ground truth in crowdsourcing. When the threshold is high enough, all inference strategies lead to ground truth when utilizing IVLs from naive attackers through label correction. In our experiments, we treat those with less than 20% correctness as a naive attacker.

VI. Summary

In this work, we use IVLs, which contain more information than binary-valued labels do, to quantitatively estimate crowd worker's reliability in terms of his/her correctness, confidence,

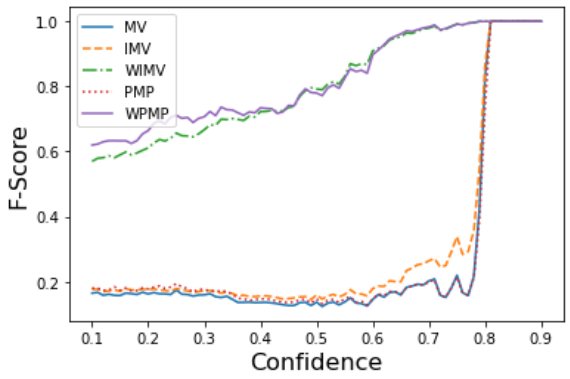

Fig. 10: F_1-score values vs. confidence threshold

stability, and predictability. Applying worker's reliability, we have developed a worker selection scheme together with two reliability weighted algorithms WIMV and WPMP to make inference in crowdsourcing. Especially when the threshold of worker's confidence is low, the newly proposed approaches lead to significant quality improvements comparing against other strategies consistently.

Our initial work reported here indicates that crowd workers' reliability may significantly impact the overall quality of crowdsourced work. We are currently applying worker's reliability, especially confidence, stability, and predictability, to anomaly detection and explainable crowdsourcing. We expect to report the results in another paper soon.

REFERENCES

[1] Bi, W., Wang, L., Kwok, J., and Tu, Z.: Learning to predict from crowdsourced data, UAI'14: Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligenc, pp. 82–91 (2014)

[2] Checco, A., Bates, J., and Demartini, G.: Adversarial attacks on crowdsourcing quality control, J. of Artificial Intelligence Research 67, pp. 375–408, (2020)

[3] Corliss, GF, Hu, C., Kearfott, RB, and Walster, GW.: Rigorous Global Search – Executive Summary, technical report, (1997)

[4] Dai, J., Wang, W., Mi, J.: Uncertainty measurement for interval-valued information systems, Information Sciences, Elsevier (2013)

[5] Dawid, A., and Skene, A.: Maximum likelihood estimation of observer error-rates using the EM algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1, pp. 20-28, (1979)

[6] Duan, Q., Hu, C., and Wei H.: Enhancing network intrusion detection systems with interval methods, SAC '05: Proceedings of the 2005 ACM symposium on Applied computing, pp. 1444–1448, (2005)

[7] Gan, Q., Yang, Q., Hu, C.: Parallel all-row preconditioned interval linear solver for nonlinear equations on multiprocessors, Parallel Computing, Vol. 20, Iss. 9, pp. 1249–1268, (1994).

[8] Goldstein, M., Uchida, S.: A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLoS ONE 11(4): e0152173, (2016)

[9] He, L., and Hu, C.: Midpoint method and accuracy of variability forecasting. J. Empirical Economics, 38, 705–715, Springer-Verlag (2009)

[10] He, L., Hu C.: Impacts of interval computing on stock market forecasting. J. of Computational Economics, 33(3), pp. 263-276, Springer (2009)

[11] Hu, C., Frolov, A., Kearfott, R., Yang, Q.: A general iterative sparse linear solver and its parallelization for interval Newton methods. Reliable Computing 1, 251—263 (1995).

[12] Hu, C., Cardenas, A., Hoogendoorn, S. et al. An interval polynomial interpolation problem and its Lagrange solution. Reliable Computing 4, 27–38 (1998).

[13] Hu, C.: Using interval function approximation to estimate uncertainty, In: Huynh VN., Nakamori Y., Ono H., Lawry J., Kreinovich V., Nguyen H.T. (eds) Interval / Probabilistic Uncertainty and Non-Classical Logics. Advances in Soft Computing, vol 46. Springer, Berlin, Heidelberg (2008).

[14] Hu, C. and *et al*: Knowledge processing with interval and soft computing. Springer-Verlag, London (2008)

[15] Hu, C. and He, L.: An application of interval methods to stock market forecasting. J. Reliable Computing, 13, pp. 423-434, Springer (2007)

[16] Hu, C.: Interval function and its linear least-squares approximation, ACM SNC '11: Proceedings of the 2011 International Workshop on Symbolic-Numeric Computation,pp. 16–23, ACM, (2012)

[17] Hu, C., and Hu ZH.: On statistics, probability, and entropy of interval-valued datasets. In: Lesot MJ. et al. (eds) Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2020. Communications in Computer and Information Science, vol 1239, pp. 422-435, Springer, Cham. (2020)

[18] Hu, C., and Hu, ZH.: A computational study on the entropy of interval-valued datasets from the stock market. In: Lesot MJ. et al. (eds) Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2020. Communications in Computer and Information Science, vol 1239, pp. 407-421, Springer, Cham. (2020)

[19] Hu, C. Sheng, SV., Wu, N., and Wu, X.: Managing uncertainties in crowdsourcing with interval-valued labeling, In: Rayz J., Raskin V., Dick S., Kreinovich V. (eds) Explainable AI and Other Applications of Fuzzy Techniques. NAFIPS 2021. Lecture Notes in Networks and Systems, vol 258, pp. 166–178, Springer, Cham. (2021)

[20] Hu, P., Dellar, M., and Hu, C.: Task scheduling on flow networks with temporal uncertainty, 2007 IEEE Symposium on Foundations of Computational Intelligence, pp. 128-135, (2007)

[21] Huynh, VN., Nakamori, Y., Hu, C. and Kreinovich, V.: On decision making under interval uncertainty: A new justification of Hurwicz optimism-pessimism approach and its use in group decision making. the 39th Int. Sym. on Multiple-Valued Logic, 214-220 (2009)

[22] Korvin, A., Hu, C., and Chen, P.: Generating and applying rules for interval valued fuzzy observations. Lecture Notes in Computer Science, vol 3177. pp. 279-284, Springer, Berlin, Heidelberg (2004)

[23] Li, H., Liu, Q.: Cheaper and Better: Selecting Good Workers for Crowdsourcing, https://arxiv.org/abs/1502.00725 (2015)

[24] Marupally, P., Paruchuri, VS., Hu, C.: Bandwidth variability prediction with rolling interval least squares (RILS). In: Proceedings of the 50th ACM SE Conference, Tuscaloosa, AL, USA, March 29-31, 2012, 209–213, ACM, (2012)

[25] Nordin, B., Hu, C., Chen, B., and Sheng, VS.: Interval-valued centroids in K-means algorithms. In: Proceedings of the 11th IEEE Int. Conf. on Machine Learning and Applications (ICMLA), pp. 478–481, Boca Raton, FL, USA, IEEE (2012)

[26] Parer, J., and Hamilton, E.: Comparison of 5 experts and computer analysis in rule-based fetal heart rate interpretation, Am J. Obstetrics Gynecology, 203(5), 451.E1–451.E7, (2010)

[27] Rajpal, S., Goel, K., and Mausam, M.: POMDP-Based Worker Pool Selection for Crowdsourcing, Proceedings of the 32nd International Conference on Machine Learning, Lille, France, (2015)

[28] Rhodes, C., Lemon, J., and Hu, C.: An interval-radial algorithm for hierarchical clustering analysis, 14th IEEE Int. Conference on Machine Learning and Applications (ICMLA), pp. 849-856, Miami, FL, USA, IEEE, (2015).

[29] Shannon, C. -E.:, A mathematical theory of communication. The Bell System Technical Journal, Vol. 27, pp. 379-423, (1948)

[30] Sheng, VS., Provost, F., and Ipeirotis, P.: Get another label? Improving data quality and data mining using multiple, noisy labelers, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 614–622, Las Vegas, Nevada, USA, August 24-27, (2008),

[31] Sheng, VS., and Zhang, J: Machine learning with crowdsourcing: A brief summary of the past research and future directions, Proc. of the 33^{rd} Conf. on Artificial Intelligence, AAAI-19, 9837–9843, (2019),

[32] Sheng, VS., Zhang, J., Bin, G., Wu, X.: Majority voting and pairing with multiple noisy labeling. IEEE Trans Knowl Data Eng 31(7):1355–1368 (2019)

[33] Qiu, L., *et al*: CrowdSelect: Increasing accuracy of crowdsourcing tasks through behavior prediction and user selection, Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, 539–548, (2016)

[34] Tao, F., Jiang, L., Li, C.: Label similarity-based weighted soft majority voting and pairing for crowdsourcing, Knowledge and Information Systems 62:2521–2538, (2020)

[35] Wang, G., Wang, T., Zheng, H., and Zhao, B.: Man vs. machine: practical adversarial detection of malicious crowdsourcing workers, Proc. of the 23rd USENIX Security Symposium, San Diego, CA, USA, August 20-22, 2014, pp239–254, USENIX Association, (2014)

[36] Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., Movellan, J.: Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise, Advances in Neural Information Processing Systems 22: 2035–2043 (2008)

[37] Wikipedia, Confusion matrix en.wikipedia.org/wiki/Confusion_matrix

[38] Zhang, J., Sheng, VS., Nicholson, B., and Wu, X.: CEKA: A Tool for Mining the Wisdom of Crowds, Journal of Machine Learning Research 16, 2853–2858 (2015).

[39] Zhang, J., Wu, X., and Sheng, VS.: Learning from crowdsourced labeled data: a survey. Artif Intell Rev 46, 543–576 (2016).