*Artificial Intelligence and Social Computing,* Vol. 72, 2023, 22–31
https://doi.org/10.54941/ahfe1003270

# Dynamically Monitoring Crowd-Worker's Reliability with Interval-Valued Labels[1]

**Chenyi Hu and Makenzie Spurling**

Computer Science and Engineering, University of Central Arkansas, Conway, AR 72035, USA

## ABSTRACT

Crowdsourcing is rapidly emerging as a computing paradigm in machine learning. Due to its open nature, human workers in crowdsourcing usually come with various levels of knowledge and socio-economic backgrounds. It has been a challenge to manage such human factors in crowdsourcing. Very recently, interval-valued labels (IVLs) have been introduced to specify worker's uncertainty (Hu et al. 2021). With IVLs, people can quantify worker's reliability, and significantly improve the overall quality of crowdsourcing (Spurling et al. 2021) and (Spurling et al. 2022). Noticing that the reliability of a worker may vary from time to time rather than a constant, we further study dynamically monitoring and updating worker's reliability in this paper.

**Keywords:** Interval-valued labels and time series, Analysing interval-valued labels, Monitoring worker's reliability dynamically, Applications of worker's reliability

## INTRODUCTION

Crowdsourcing has become a popular paradigm in machine learning. It gathers labels from large groups of people (crowd-workers), usually through the internet.

The quality of crowdsourcing very much depends on the quality of the collected labels in addition to learning algorithms (Sheng and Zhang, 2019). Due to its open nature, human factors are unavoidably involved in crowdsourcing. Barbosa and Chen reported that human biases caused by worker's social-economic status can determine the outcome of crowdsourcing tasks (Barbosa and Chen, 2019). Another issue is that the level of expertise for crowd-workers often varies widely. Workers with a high level of expertise usually give better quality labels than those with less. However, a highly knowledgeable worker with adversarial intentions can cause more harm. Properly managing human factors in crowdsourcing has been actively studied. For instance, Bi et al studied the impacts of worker's dedication, expertise, judgment, and task difficulty in (Bi et al. 2014). Qiu et al offered methods for selecting workers based on behavior prediction (Qiu et al. 2016). Tao et al reported quality improvements in (Tao et al. 2020) when utilizing MV-Freq and MV-Beta (Sheng et al. 2019) with worker's reliability.

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

 

In contrast to commonly used binary-valued labels in previous work, Hu et al proposed interval-valued labels (IVLs) recently (Hu et al. 2021). Applying statistical and probabilistic properties of interval-valued datasets, one can quantify worker's reliability and achieve significant quality improvement (Spurling et al., 2021) and (Spurling et al., 2022). Worker's behavior in real world often varies from time to time and is not consistent as implicitly assumed in previous study. We should monitor and update workers' reliability dynamically.

## A BRIEF REVIEW OF PREVIOUS WORK

Prior to our discussion, let us briefly review related concepts, notations, and previous results as background knowledge first.

### Statistic and Probabilistic Properties of an Interval-Valued Dataset

In this paper, we use interval-valued labels. Following the literature of interval computing, we use boldface letters to separate interval-valued objects from point-valued ones. For example, $a$ is a real where $\boldsymbol{a}$ is an interval. The greatest lower and least upper bounds of an interval are specified with an underline and an overline on the same non-boldface letter. Thus, $\boldsymbol{a} = [\underline{a},\ \overline{a}]$. The midpoint and radius of $\boldsymbol{a}$ are $\text{mid}(\boldsymbol{a}) = (\underline{a} + \overline{a})/2$ and $\text{rad}(\boldsymbol{a}) = (\overline{a} - \underline{a})/2$. As the two are point-valued, we write them without boldface as $\text{mid}(a)$ and $\text{rad}(a)$ hereafter.

Let $\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_m]$ be a list of intervals. The midpoint and radius of $\boldsymbol{L}$ are $\text{mid}(L)$ and $\text{rad}(L)$, respectively. The mean of $\boldsymbol{L}$ is the interval:

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

The variance of $\boldsymbol{L}$ defined is a real as

$$\text{Var}(L) = \text{Var}\left(\text{mid}(L)\right) + \text{Var}\left(\text{rad}(L)\right) + \frac{2}{m}\sum_{i=1}^{m}|\Delta m_i \Delta r_i| \tag{2}$$

where $\Delta m_i = \text{mid}(\boldsymbol{l}_i) - \mu\left(\text{mid}(L)\right)$ and $\Delta r_i = \text{rad}(\boldsymbol{l}_i) - \mu\left(\text{rad}(L)\right)$. So, we can calculate the standard deviation of $\boldsymbol{L}$ as usual:

$$\sigma(L) = \sqrt{\text{Var}(L)}. \tag{3}$$

Hu and Hu (2020a) also defined a probability density function (pdf) for $\boldsymbol{L}$ as

$$f(t) = \frac{\sum_{i=1}^{m} pdf_i(t)}{m}, \tag{4}$$

where $\text{pdf}_i(t)$ is the pdf of an $\boldsymbol{l}_i \in \boldsymbol{L}$. With the pdf, we can calculate Shannon's entropy of $\boldsymbol{L}$.

### Reliability Indicators Derived from Crowd-Worker's IVLs

We assume a binary classification model in this work. Let $V$ be a set of observations. The objective of a crowdsourced task is to determine whether a $v_i \in V$ is an instance of a given class $y$ or not with multi-labels provided by some $j \in J$, which is the set of crowd-workers. We use $\boldsymbol{l}_i^j$ to denote the IVL by a worker $j \in J$ for a given $v_i \in V$. Because of the binary classification model, $\boldsymbol{l}_i^j = \left[\underline{l}_i^j, \overline{l}_i^j\right] \subseteq [0, 1]$. The IVL contains j's uncertainty on $v_i$. A real is in fact a narrow interval with its greatest lower and least upper bounds the same. IVLs extend binary-valued labels 0 and 1.

Let $\boldsymbol{L}^j$ be the list of IVLs made by a worker $j \in J$. From which, one may quantify $j$'s reliability into four reliability indicators: correctness, confidence, stability, and predictability (Spurling et al., 2021). The *correctness* of a worker is the ratio of accurately labeled observations by the worker. A set of gold questions with known ground truth is commonly used to estimate a worker's correctness. Let $G = g$ be a set of gold questions. The IVL from a worker $j$ on a $g \in G$ is $\boldsymbol{l}_g^j$. Denoting the ground truth of $g$ as $o(g)$. Then, the center-correctness of $\boldsymbol{l}_g^j$ is

$$\begin{cases} 1 - \text{mid}\left(\boldsymbol{l}_g^j\right) & \text{if } o(g) = 0 \\ \text{mid}\left(\boldsymbol{l}_g^j\right) & \text{if } o(g) = 1 \end{cases}.$$

Without loss of generality, we can assume the ground truth is 1 for any $g \in G$ in calculating $j$'s correctness. This is because of that, when $o(g) = 0$, we can replace $\boldsymbol{l}_g^j$ with its difference from 1, i.e. $1 - \boldsymbol{l}_g^j = \left[1 - \overline{l}_g^j, 1 - \underline{l}_g^j\right]$ without changing the center-correctness. Let $\boldsymbol{L}_G^j$ be $j$'s IVLs on the set of gold questions $G$. We can apply Eq. (1) to calculate the mean of $\boldsymbol{L}_G^j$, $\mu(\boldsymbol{L}_G^j)$, which estimates $j$'s average correctness.

An IVL $\boldsymbol{l}_i^j$ contains information of $j$'s *confidence* too. When $\text{mid}\left(\boldsymbol{l}_i^j\right) = 0.5$, $j$ has absolutely no confidence toward either 0 or 1. Otherwise, $\text{mid}\left(\boldsymbol{l}_i^j\right)$ represents $j$'s preference toward 0 or 1. The distance between the midpoint and 0.5, i.e., $\left|\text{mid}\left(\boldsymbol{l}_i^j\right) - 0.5\right|$, reflects confidence of $\boldsymbol{l}_i^j$. Additionally, the radius of $\boldsymbol{l}_i^j$ specifies the maximum possible variation from the midpoint. When $\text{rad}\left(\boldsymbol{l}_i^j\right) = 0$, the label is point-valued. Otherwise, the label contains $j$'s uncertainty over a range. The maximum possible value of $\text{rad}\left(\boldsymbol{l}_i^j\right)$ is 0.5. The difference between 0.5 and $\text{rad}\left(\boldsymbol{l}_i^j\right)$, $0.5 - rad\left(\boldsymbol{l}_i^j\right)$, is another measure of confidence. Ultimately, the confidence of a single $\boldsymbol{l}_i^j$ is a combination of the above two. That is:

$$\text{conf}\left(\boldsymbol{l}_i^j\right) = \left|\text{mid}\left(\boldsymbol{l}_i^j\right) - 0.5\right| + 0.5 - rad\left(\boldsymbol{l}_i^j\right).$$

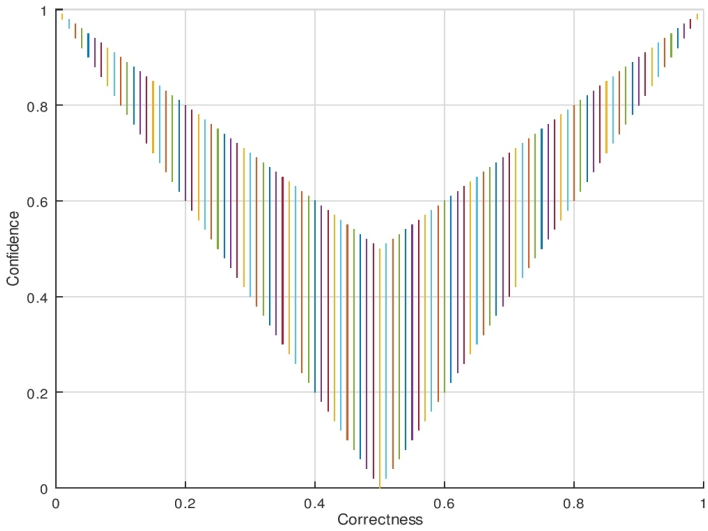


**Figure 1:** The relation between worker's correctness and confidence (Spurling et al. 2021).

Since both $\left|\text{mid}\left(l_i^j\right) - 0.5\right|$ and $0.5 - \text{rad}\left(l_i^j\right)$ are between 0 and 0.5, the confidence of $l_i^j$ can be any real between 0 and 1. It is important to notice that the confidence of an IVL does not depend on the ground truth. Thus, the mean of $\boldsymbol{L^j}$ reflects $j$'s overall confidence as

$$\left|\text{mid}\left(\mu\left(L^j\right)\right) - 0.5\right| \; + \; 0.5 - \text{rad}\left(\mu\left(L^j\right)\right). \tag{5}$$

Fig. 1 illustrates the relationship between worker's correctness and confidence (Spurling et al. 2021), i.e., one's correctness and confidence should be a point inside the V shaped area.

The overall *stability* of $j$ is reflected in $\sigma\left(L^j\right)$, which can be calculated with Eqs. (2) and (3). In addition, Eq. (4) provides a pdf of $\boldsymbol{L^j}$. Hence, we can calculate Shannon's entropy of $\boldsymbol{L^j}$ as an estimation of $j$'s *predictability*. It is important to notice that estimating $j$'s confidence, calculating $j$'s stability and predictability does not require the ground truth of a $v \in V$.

## DYNAMICALLY MONITORING WORKER'S RELIABILITY

In this section, we discuss the needs and approaches to monitor worker's reliability dynamically.

### We Need to Monitor Worker's Reliability Dynamically

Applying the quantified reliability measures above, we can effectively take human factors into consideration. For instance, reliability weighted inference making schemes such as weighted interval majority voting (WIVM) and weighted preferred matching probability (WPMP) can significantly improve

overall quality (Spurling et al., 2021). A challenging task in crowdsourcing is to identify and exclude labels from those who are not reliable even with adversary purposes. In fact, attackers may pretend to be regular workers with the sole purpose to derail a crowdsourced task (Checco et al., 2020, Wang et al. 2014). Very sophisticated attackers are likely to identify gold questions and answer them with a high level of correctness, and deliberately label regular questions incorrectly (Qiu et al. 2016). Monitoring worker's reliability derived from IVLs, one may identify such anomalous workers.

Let $G$ be the set of gold questions and $U$ be the set of regular questions. We can partition $\boldsymbol{L}^j$ as $\boldsymbol{L}^j_G$ and $\boldsymbol{L}^j_U$. Assuming $G$ well samples $U$. Then, $\boldsymbol{L}^j_G$ and $\boldsymbol{L}^j_U$ should be consistent statistically when $j$ behaves normally. Otherwise, it implies possible anomaly. Applying the t-test and F-test, we can examine the consistency between $\boldsymbol{L}^j_G$ and $\boldsymbol{L}^j_U$ statistically without the ground truth of any u $\in U$. With this approach, Spurling et al. (2022) successfully identified attackers and achieved significant quality improvements. However, it was implicitly assumed that the reliability of a given worker is constant. This is not generally true in real world applications. For instance, a worker with malicious intention may purposely change his/her labelling pattern from time to time. A dedicated worker may improve his/her reliability along with more experience/knowledge. Moreover, random factors such as changing working environment, emotion, stress level, and others can cause variations of worker's reliability. Therefore, we need to monitor and update worker's reliability dynamically.

### Ways of Monitoring and Updating Worker's Reliability Dynamically

The reliability of a worker $j$ varies from time to time. Thus, $j$'s reliability is a function of time $r^j(t)$. The reliability of $j$ is derived from the list of his/her IVLs $\boldsymbol{L}^j$. Because $j$ labels $v_i \in V$ sequentially, $\boldsymbol{L}^j = \left[\boldsymbol{l}^j_1, \boldsymbol{l}^j_2, \cdots, \boldsymbol{l}^j_k, \boldsymbol{l}^j_{k+1}, \cdots\right]$ is an interval-valued time series. To monitor worker's behavior dynamically, we assume that the reliability value of $j$, $r^j$, at time $t$ depends on some consecutive IVLs in $\boldsymbol{L}^j$ within a time window $T$. Assuming the first $k$ IVLs in $\boldsymbol{L}^j$ are in the window $T_0$. We can then initiate $j$'s reliability at $t = t_0$ as $r^j(t_0) = f\left(\boldsymbol{l}^j_1, \boldsymbol{l}^j_2, \cdots, \boldsymbol{l}^j_k\right)$ in terms of correctness, confidence, stability, and predictability. As discussed earlier, the correctness of $j$'s reliability relies on $j$'s IVLs on gold questions. To monitor $j$'s correctness dynamically, we need to present gold questions periodically. When the IVL on a new gold question from $j$ becomes available, we can re-evaluate $j$'s correctness and other reliability indicators. Let $\boldsymbol{l}^j_g$, be the IVL for a new $g \in G$ at $t_0 < t \le t_1$ by a worker $j$. To update $j$'s reliability $r^j(t_1)$, we use IVLs in the time window $T_1$ only, which include the newly available $\boldsymbol{l}^j_g$. Continuing the process, we can dynamically evaluate $j$'s current reliability in a moving time window $T_i$. (Note: worker's inactive time should be excluded.) One should apply worker's current reliability in inference making. For anomaly detection, one needs to examine and compare the reliability derived from the IVLs on gold questions $\boldsymbol{L}^j_{G_{T_i}}$ and from those on regular questions $\boldsymbol{L}^j_{U_{T_i}}$ within the time window $T_i$.

As the time window moves forward, the worker $j$'s correctness, confidence, stability, and predictability are updated. Each of these four reliability indicators forms a time series too. With a user pre-selected orthogonal basis of function space, say $f_0, f_1, \ldots, f_s$, one may fit $j$'s reliability as a linear combination of the basis, i.e., $r^j(t) \approx \sum_{i=0}^{s} \beta_i f_i(t)$, through regression analysis. The approximated $r^j(t)$ can separate the observed variations of $j$'s reliability into an explainable trend together with an error term. In studying variations of $j$'s reliability, it may suggest possible anomaly when the unexplainable error is far away from expected values. For regression analysis of interval-valued data and applications, readers may refer (Hu 2007, Hu 2008, Hu 2012, Hu and Hu 2020b).

## COMPUTATIONAL EXPERIMENTS

In our computational experiments, we investigate the impact of worker's reliability on the overall quality of crowdsourced work.

### Datasets, Methods, and Quality Measures

In our experiments, we use four binary-classification benchmark datasets named Car, Income94, Sick and Vote, with known ground truth, in CEKA (Zhang et al. 2015) as testing datasets. Table 1 lists the size, number of attributes, and number of positive and negative instances of each dataset. The datasets are imbalanced except Income94. We use a portion of the test dataset as gold questions, and the rest as regular questions.

A virtual pool of one hundred workers is generated with various levels of reliability. Five inference making schemes are employed in our experiments. Three of them serve as baseline schemes. They are majority voting (MV), interval majority voting (IMV), and preferred matching probability (PMP). Two reliability weighted schemes are weighted interval majority voting (WIMV) and weighted preferred matching probability (WPMP). Because the ground truth of each observation is known, we can form the confusion matrix in our experiment. We measure the overall quality with accuracy, precision, recall, and $F_1$-score derived from confusion matrix.

### Reliability Weighted Schemes Improve the Overall Quality

Experiments on all four test datasets demonstrate that the reliability weighted schemes can significantly improve the overall quality of crowd-sourced tasks. Figure 2 illustrates recall, precision, accuracy, and $F_1$-score on the dataset Car with the five inference making schemes. In which, the horizontal

**Table 1.** Datasets used in experiments.

| Name | Size | Attributes | Positive | Negative |
|---|---|---|---|---|
| Car | 1,594 | 7 | 384 | 1,210 |
| Income94 | 600 | 15 | 300 | 300 |
| Sick | 3,773 | 30 | 231 | 3,541 |
| Vote | 435 | 17 | 168 | 267 |

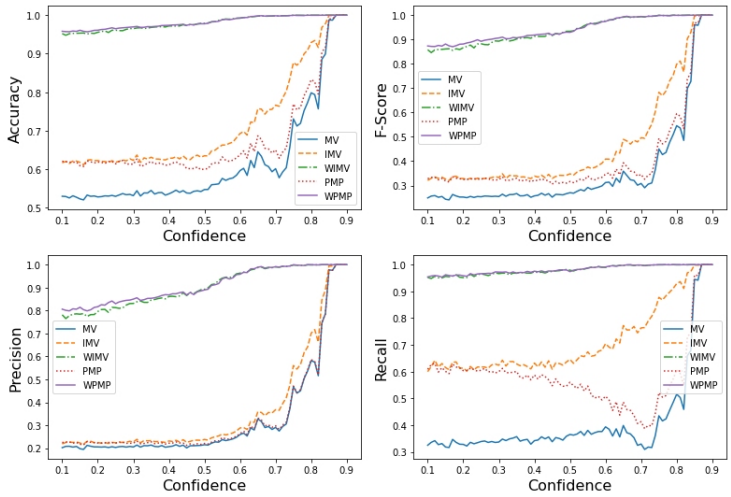


**Figure 2:** Comparison of performance on the test dataset car.

axis indicates the minimum confidence threshold for worker selection. Above each confidence threshold, ten workers are randomly selected from the pool for each run. The workers give IVLs on all questions; and the performance metrics are calculated based on the resulting confusion matrices. These scores are then averaged over forty runs to minimize any outliers due to the random factor. To utilize IVLs from workers with a low correctness ($\leq 20\%$) and a high level of confidence ($\geq 80\%$), we replace each of them with its difference from 1. The reliability weighted schemes WIMV and WPMP significantly outperform the three baseline schemes. Similar results are observed on the other three test datasets too. We do not include them in this paper to meet the page limit.

## Applying Worker's Reliability to Detect Anomaly

Dynamically monitoring worker's reliability can help us to detect anomaly and identify possible attackers. In our experiments, we set a fifth of the worker pool to

behave abnormally as designated as anomalous workers. Their labeling patterns on $U$ and $G$ are inconsistent. Applying the $t$- and $F$-tests for each $j \in J$ on $\boldsymbol{L}_{\boldsymbol{G}}^{j}$ and $\boldsymbol{L}_{\boldsymbol{G}}^{j}$, we can detect them as illustrated in Fig. 3. In which, each worker is specified with his/her correctness and confidence as a light blue dot. A dark blue x indicates an identified anomalous worker. The probabilistic confidence level is 95% in the experiment.

## Excluding Identified Attackers for Quality Crowdsourcing

In our experiments, we further investigate the impact of excluding identified attackers. Fig. 4 compares the $F_1$-score for each test dataset with or without excluding labels from identified attackers. The horizontal axis indicates the number of attackers present in the entire pool. By excluding identified attackers, we have kept a near perfect $F_1$-score for all test datasets. In contrast,

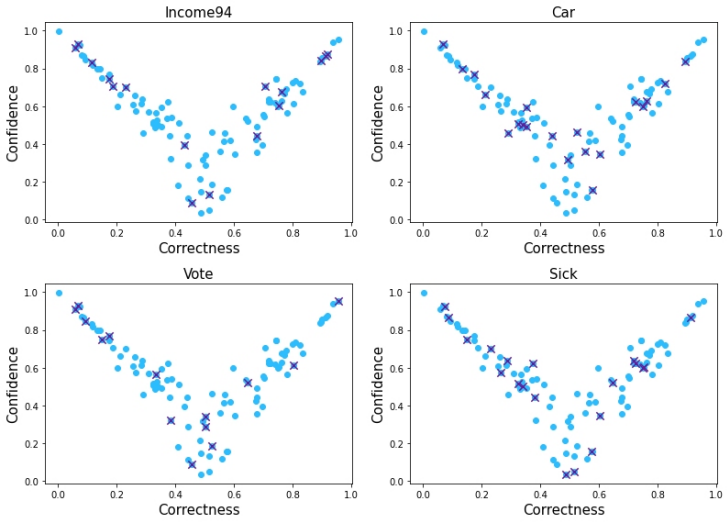


**Figure 3:** Detected possible attackers in labelling each dataset.

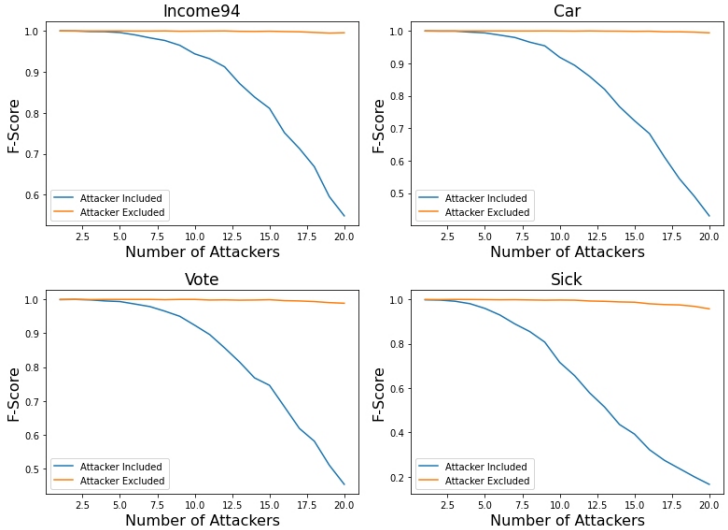


**Figure 4:** Comparisons $F_1$-score with or without excluding identified attackers.

the $F_1$-score decreases rapidly as the number of attackers increases without attacker exclusion.

## Impacts of Window Size on Monitoring Worker's Reliability

Monitoring $j$'s reliability dynamically, we should use a moving time window. Because a worker may not label instances all the time, we use a fixed number of IVLs in a window as 5, 10, 50, 100, 150, and 200 in our experiment instead. We run the $t$- and F-tests for $\boldsymbol{L}_G^j$ and $\boldsymbol{L}_U^j$. Fig. 5 illustrates the impact

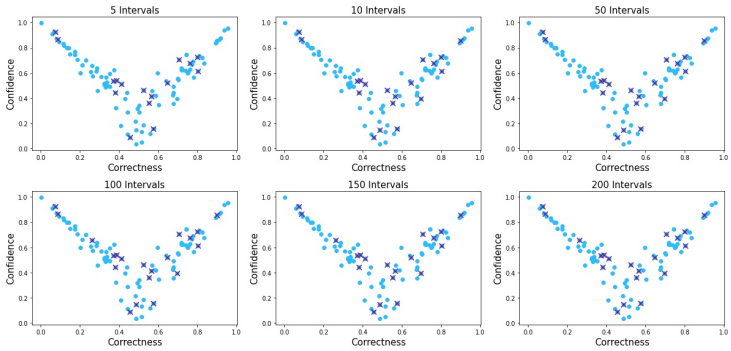


**Figure 5:** Impacts of window size on monitoring worker's reliability.

of the number of IVLs in a window on the test dataset Car[1]. As the number of IVLs increases, the anomalies start to be singled out more. This is because the system remembers actions from previous IVLs. This is fine. We need enough samples in statistic and probabilistic tests.

## CONCLUSION

Due to the open nature of crowdsourcing, crowd workers usually come with various levels of knowledge, social-economic backgrounds, and motivations. The reliability of crowd workers can impact the quality of crowdsourcing significantly. Using IVLs instead of common binary-valued ones, we can quantify the reliability of a particular worker in terms of correctness, confidence, stability, and predictability. Our computational experiments have demonstrated that applying worker's reliability, one can improve the overall quality of crowdsourcing through reliability weighted inference making, anomaly detection, and attacker exclution. A worker's reliability often varies from time to time rather than constant. We need to monitor worker's behavior and update worker's reliability dynamically in practice. In this work, we treat IVLs from a worker $j$ as a time series. Using IVLs within a forward moving time window, we can update and monitor $j$'s reliability dynamically. We can further analyze worker's reliability with regression analysis of interval-valued sequence. Results of computational experiments indicate that we need, and we can monitor worker's reliability dynamically.

## ACKNOWLEDGMENT

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

## REFERENCES

Barbosa, N., and Chen, M. (2019) "Rehumanized crowdsourcing: a labeling framework addressing bias and ethics in machine learning". Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-12.

[1]Similar results are observed on other test datasets too.

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

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

Hu, C. (2008). Using interval function approximation to estimate uncertainty. Advances in Soft Computing, vol 46. Springer, Berlin.

Hu, C. and He, L. (2007). An application of interval methods to stock market forecasting. J. Reliable Computing 13, pp. 423–434.

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

Hu, C., and Hu ZH. (2020a). On statistics, probability, and entropy of interval-valued datasets. Communications in Computer and Information Science 1239, pp. 407–421.

Hu, C., and Hu ZH. (2020b). A Computational Study on the Entropy of Interval-Valued Datasets from the Stock Market. Communications in Computer and Information Science, vol. 1239, pp. 422–435.

Hu, C. Sheng, VS., Wu, N., and Wu, X. (2021). Managing uncertainties in crowdsourcing with interval-valued labeling. Lecture Notes in Networks and Systems 258, pp. 166–178.

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

Sheng, VS, and Zhang, J. (2019). Machine learning with crowdsourcing: A brief summary of past research and future directions. Proceedings of the AAAI Conference on Artificial Intelligence 33(1), pp. 9837–9843.

Sheng, VS., Zhang, J., Bin, G., Wu, X. (2019) Majority voting and pairing with multiple noisy labeling. IEEE Trans. on Knowledge and Data Eng. 31(7), pp. 1355–1368.

Spurling M., Hu C., Zhan, H., Sheng VS. (2021). Estimating crowd-worker's reliability with interval-valued labels to improve the quality of crowdsourced work. 2021 IEEE (SSCI), pp. 01–08.

Spurling M., Hu C., Zhan, H., Sheng VS. (2022). Anomaly detection in crowdsourced work with interval-valued labels. Communications in Computer and Information Science, vol 1601. Springer, Cham.

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

Wang, G., Wang, T., Zheng, H., and Zhao, B. (2014). Man vs. machine: practical adversarial detection of malicious crowdsourcing workers. Proc. of the 23rd USENIX Security Symposium, pp. 239–254.

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