How Clumpy is my Image? Evaluating Crowdsourced Annotation Tasks

Hugo Hutt,* Richard Everson,* Murray Grant,† John Love,† George Littlejohn†
* Computer Science, † Biosciences,
The University of Exeter,
Exeter, UK.
Email: {hwh202, R.M.Everson, M.R.Grant, J.Love, G.R.Littlejohn}@exeter.ac.uk

Abstract—The use of citizen science to obtain annotations from multiple annotators has been shown to be an effective method for annotating datasets in which computational methods alone are not feasible. The way in which the annotations are obtained is an important consideration which affects the quality of the resulting consensus estimates. In this paper, we examine three separate approaches to obtaining scores for instances rather than merely classifications. To obtain a consensus score annotators were asked to make annotations in one of three paradigms: classification, scoring and ranking. A web-based citizen science experiment is described which implements the three approaches as crowdsourced annotation tasks. The tasks are evaluated in relation to the accuracy and agreement among the participants using both simulated and real-world data from the experiment. The results show a clear difference in performance between the three tasks, with the ranking task obtaining the highest accuracy and agreement among the participants. We show how a simple evolutionary optimiser may be used to improve the performance by reweighting the importance of annotators.

I. INTRODUCTION

The increasing digitisation of science has dramatically reduced the cost of generating large amounts of data in a relatively short amount of time. However, the analysis and classification of this data often requires human participation to arrive at an accurate consensus. In addition, annotating large datasets has become one of the major bottlenecks for developing effective machine learning models which can generate new predictions [1]. Alternatives to purely computational approaches are therefore required in order to obtain the annotations.

Citizen science seeks to elicit the help of non-experts to address scientific problems by using crowdsourcing. This can take the form of an online annotation task in which the collective efforts of many individual participants are used to arrive at estimates of the consensus annotations. Recently, a number of citizen science projects have shown effectiveness in using crowdsourcing approaches to acquire annotations from multiple annotators, generating datasets which can then be used to guide computational approaches [2]–[4]. Annotations gathered from citizen science experiments can result in valuable training data for machine learning models, while also providing insights into the behaviour of the participants. In addition, there are a number of interesting theoretical problems surrounding citizen science as a result of the different degrees of accuracy associated with the participants and the uncertainty inherent in the data.

With the increasing number of online projects, there is a corresponding need to investigate how crowdsourcing tasks should be presented to the participants [4]–[6]. The effect that different types of annotation tasks have on the performance and consensus of the participants is an important, but largely unexplored topic. The choice of task is an essential consideration when using crowdsourcing to gather annotations, as it determines to a significant extent the quality of the resulting data.

The goal of this paper is to investigate a number of separate approaches to obtaining annotations from experimental participants and to examine their effectiveness. We describe a web-based citizen science experiment involving the annotation of microscopy images of plant cells during bacterial infection. Briefly, the goal was to assess the degree of “clumpiness” of each image. This task therefore differs from the more common classification task, in which the annotator is asked to assign an object to discrete categories, because “clumpiness” is a continuous quantity. Three separate paradigms were used to obtain the image annotations and in this paper we assess their efficacy. The approaches are evaluated on both simulated and real-world data from the experiment and a comparison is made between the different tasks. In particular, the influence of the task type on the overall performance and consensus of the participants is examined. The annotation of the microscopy images is a very challenging problem for current image processing techniques, which makes it a good candidate for a citizen science project.

The rest of the paper is organised as follows. Section II is a description of the problem. Section III describes the citizen science experiment, including the user statistics for each of the tasks. Section IV outlines methods for evaluating different annotation tasks. Section V describes the simulation setup used to model annotators under the different tasks. Section VI presents the empirical results from the simulated and experimental data. Section VII describes how the estimates of individual annotators can be reweighted using an evolutionary optimiser to obtain more accurate results. Section VIII concludes the paper.
II. DESCRIPTION OF PROBLEM

A. Learning from Multiple Annotators

In a typical annotation task, there is a set of $N$ instances $x = \{x_1, \ldots, x_N\}$ whose true annotations are unknown. Each instance $x_i \in x$ is then assigned an annotation by $R$ annotators, resulting in a set of estimates $\{y_i^1, \ldots, y_i^R\}$ of the true annotation. Given these multiple annotations, the goal is to arrive at accurate consensus estimates $y = \{y_1, \ldots, y_N\}$ for each of the $N$ instances.

One simple and often used technique for obtaining consensus estimates from multiple annotators is majority voting [7]. For binary classification, the majority vote estimate of an instance $x_i$ is defined as

$$y_i = \begin{cases} 1 & \text{if } \frac{1}{R} \sum_{j=1}^{R} y_i^j \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

(1)

where $y_i^j \in \{0, 1\}$ is the annotation assigned to instance $x_i$ by annotator $j$. For simplicity of notation we assume that each instance is annotated by the same number, $R$, of annotators, although in practice $R$ is often different for each instance. Majority voting can be extended to scores, where each instance is assigned the mean of the annotators’ scores:

$$y_i = \frac{1}{R} \sum_{j=1}^{R} y_i^j$$

(2)

If the scores are made on an integer scale (e.g., a five-point scale: $y_i^j \in \{1, 2, 3, 4, 5\}$), the estimate $y_i$ can then be rounded to the nearest score on the scale.

Ideally, annotators with higher accuracy should be given more weight when estimating the consensus, while the influence of poor quality annotators should be decreased or removed entirely. A major limitation of standard majority voting is that it assumes all annotators are equally reliable, meaning that its effectiveness is dependent on the overall quality of the annotators [7]. Given an estimate of an annotator’s performance, we can introduce an additional weighting term to the vote to account for the variation in quality among the annotators. Let $\epsilon_j$ be the error rate of annotator $j$ on some subset of the instances for which the true annotations are known. The standard majority vote can be replaced by

$$y_i = \begin{cases} 1 & \text{if } \frac{1}{R} \sum_{j=1}^{R} |y_i^j - \epsilon_j| \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

(3)

which penalises annotators with high error rates and assigns greater weight to the estimates of accurate annotators. This approach can be extended to scores, for example by weighting by the mean signed error. If there is no known standard on which to evaluate the annotators, more sophisticated techniques are required in order to account for the differences in annotator quality, such as those proposed in [1], [2], [7].

In Section VII we describe how annotator weighting parameters can be optimised to improve the accuracy of the estimates.

B. Obtaining Scores from Annotators

The annotations collected from the experiment indicate the degree of “clumpiness” present in the microscopy images. This notion of clumpiness is continuous in nature, with scores potentially falling within an indefinite range. Unlike classification tasks, which involve assignment to predefined categories, ways of assessing a score are less well explored. We therefore asked annotators to perform three different kinds of task in order to elicit a consensus score. The following is a description of the three kinds of annotation investigated in this paper: classification, scoring and ranking.

The classification task divides the range of scores into two (not clumpy and clumpy) and requires the annotators to assign binary scores $\{0, 1\}$ to the instances:

$$y_i^j \in \{0, 1\} \ \forall x_i \in x$$

(4)

A consensus classification is then obtained by majority voting (1). In addition, a score is obtained as the proportion of 1 annotations assigned (eq. (2)). Clearly, this score can be interpreted as the probability that the instance belongs to either class. To obtain the maximum amount of information the class boundary should be placed so that roughly half of the instances fall in either class. However, while this task is conceptually straightforward for an annotator, they may find it difficult to assign instances close to the artificially-imposed division between the classes. Furthermore, the extreme “quantisation” of the continuous scale into just two categories inevitably discards information about degree which is only recovered after many annotations have been made.

For the scoring task, the annotators directly assign scores in a pre-determined range. Although in principle an indefinitely fine scale could be employed, in our experiments a seven-point integer scale was used:

$$y_i^j \in \{1, 2, 3, 4, 5, 6, 7\} \ \forall x_i \in x$$

(5)

A fairly coarse integer scale, like this, relieves annotators of feeling that they have to make very fine distinctions, while allowing them to distinguish between very clumpy and quite clumpy, etc. Nonetheless, even when furnished with examples, annotators may not use the full range of the scale and, of course, may assign different scores based on their prejudices and the particular instances that they have seen previously. Clearly, a consensus score is easily given by the mean of the annotators’ scores (2).

For the ranking task, annotators are required to rank-order subsets of the instances according to whatever quantity is being assessed. This results in a set of ordered relations and we write $(x_i \prec_j x_k)$ to indicate that $x_i$ has been assessed to have a lower score than $x_k$ by annotator $j$. Consider the specific case in which each instance is ranked either higher or lower than one other instance. From these binary rankings, a score is derived for each of the instances. Let

$$r_{x_i} = \{x_k \in x \mid (x_k \prec_j x_i) \forall j\}$$

(6)

be the set of instances ranked lower than $x_i$ by any annotator. Also let

$$t_{x_i} = \{x_k \in x \mid (x_k \prec_j x_i) \lor (x_i \prec_j x_k) \forall j\}$$

(7)
be the instances ranked either lower or higher than \( x_i \). The consensus score for \( x_i \) is then
\[
y_i = \frac{|r_{x_i}|}{|t_{x_i}|}
\]  
(8)

Clearly, instances that are consistently ranked above other instances will obtain high scores and vice versa. The advantage of the ranking task is that annotators find it relatively easy to compare instances and agree on an ordering even if they disagree on a precise score or even to which of a pair of classes an instance belongs. Unlike the classification and scoring tasks there is no need for the annotator to refer back to a set of fiducial instances for calibration.

III. DESCRIPTION OF EXPERIMENT

In order to assess the different approaches to obtaining the instance annotations outlined above, we describe here a web-based citizen science experiment involving the annotation of plant cell images according to their “clumpiness”. 1

The microscopy images obtained for the experiment show perfluorocarbon-mounted [8] leaves of the model plant Arabidopsis thaliana (Col-0 ecotype) obtained using a Zeiss 510Meta Laser Scanning Confocal Microscope equipped with a 40x oil immersion lens. Chlorophyll was imaged by Excitation at 488 nm and Emitted Fluorescence was collected with a LP615 nm filter. Z-stacks, consisting of 75 \( \times \) \(-\)steps were collected during a timecourse comparing infection with the phytopathogenic bacterium Pseudomonas syringae pv. tomato strain DC3000 to a mock inoculation using an actived at the time of writing.

1The URL of the experiment is http://www.clumpy.ex.ac.uk which remains active at the time of writing.
Fig. 1: Examples of the chloroplast images used for the experiment. Also shown are the consensus scores from the ranking task, which range from 0 to 1.

Fig. 2: The good and adversarial annotators are above and below the diagonal of the ROC, respectively. Spammers are close to the diagonal of the ROC, assigning the scores at random.

Annotations to the instances. The proportion of good annotators in the population and their overall level of performance depends on a number of factors, such as the individual difficulty of the instances, the duration of the task, as well as the ability of the annotators.

An adversarial annotator’s curve lies below the diagonal of the ROC. These annotators are the mirror image of the good annotators on the ROC plot, assigning incorrect annotations to the instances. An important point to note is that although adversarial annotators are inaccurate and assign incorrect annotations, they do so consistently. This means that if they can be detected in the population and have their annotations “flipped”, they still have discriminatory power [1].

Finally, a spammer can be viewed as an annotator who assigns annotations at random [1]. For binary classification, this corresponds to the situation in which the annotator is close to the diagonal of the ROC plot, as shown in Figure 2. Annotators close to the diagonal of the ROC provide no useful discriminatory power and their annotations should be ignored or removed if they are detected in the population.

Although annotators tend to fall into one of the three classes, the distinction is not always easy to make. For example, an annotator may start off as random or adversarial, but improve their accuracy as they are exposed to more instances. Conversely, an annotator’s accuracy can also decrease over time.

In addition to evaluating the accuracy, a number of other properties of the annotators were assessed when comparing the three tasks. We measured how strongly the annotators were correlated with each other and how reliable they were in maintaining their accuracy for the duration of the task. The results are presented in Section VI.

Annotations in the ranking class are easily cast in a classification framework by considering the binary relations that result from ordering the instances. We denote a ranking as correct if the two instances involved are placed in their true order (obtained from the gold standard or consensus) and incorrect if not.

Finally, the scoring task is cast in a classification framework denoting a score as correct if it and the true score are both greater than or equal to 4 (the middle of the available range) or if both are less than or equal to 4; otherwise the score is deemed incorrect. Although other more sensitive loss functions might be used in this context, this provides a common framework for evaluating performance in all three tasks.

V. SIMULATION

In order to investigate the annotation tasks under various conditions, simulated data was generated to model annotators with different degrees of accuracy and performance. In a simple model, whether each annotator correctly classifies an instance could be modelled by a draw from a Bernoulli distribution $\text{Be}(\pi)$ with an annotator-specific probability $\pi$. To provide a richer model, the probability $\pi$ was modelled by a beta distribution. The beta distribution is defined by

$$p(\pi; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1}$$

where the two parameters $\alpha > 0$ and $\beta > 0$ control the mean and spread of the distribution. This is demonstrated in Figure 3, which shows the beta distribution for different values.
of the parameters. The values of $\alpha$ and $\beta$ were assumed to be annotator specific.

VI. EMPIRICAL RESULTS

A. Accuracy of Annotators

Estimates of the annotator parameters for the beta distribution were obtained using maximum likelihood and the observed accuracy on the gold standard. The beta distributions using these parameter estimates are shown for each of the tasks in Figure 4. The ranking task distributions tended to be more sharply peaked compared to the other two tasks, with $\mu = 0.75$ and $\sigma = 0.09$ for the mean distribution. This indicates that there was less variation in accuracy among the annotators. The participants from the classification ($\mu = 0.75$, $\sigma = 0.15$) and scoring ($\mu = 0.68$, $\sigma = 0.17$) tasks tended to be less reliable in their estimates, obtaining a wider range of accuracies.

Using the parameter estimates, the ROC curves of simulated annotators were obtained for each of the tasks. Figure 5 shows the results. The ROC curves derived from the mean parameter estimates are shown in bold. A number of annotators were close to the diagonal of the ROC, indicating the presence of spammers in the population. Adversarial annotators can also be clearly identified in each of the tasks, as shown by the curves lying below the diagonal of the ROC. The mean curves are seen to obtain good performance, with the ranking task in particular being near-optimal.

The accuracy of the annotators was also evaluated in relation to the consensus (majority vote) annotations for the images. Figure 6 shows the consensus accuracy for annotators versus the number of image annotations they made. As can be seen, the consensus accuracy of the annotators tended to remain stable, with no large increase or decrease in the accuracy as more annotations were made.

In terms of the overall accuracy, the ranking task obtained the best performance. The greater proportion of annotators with high accuracy was reflected in the performance of the majority vote estimates.

B. Inter-Annotator Agreement

The inter-annotator agreement provides a measure of the consensus among multiple annotators, which enables a comparison between different annotation tasks in terms of the agreement among the participants. We used the Spearman rank correlation for the comparison, which is a non-parametric statistic measuring the strength of association between two sets of data. Let $\{y^i_j\}$ and $\{y^k_j\}$, $i = 1, \ldots, R$, be the sets of image annotations in common between annotators $j$ and $k$. The Spearman correlation between the two annotators can then be defined as

$$\rho_{jk} = 1 - \frac{6 \sum_{i=1}^{R} (\sigma^i_j - \sigma^k_j)^2}{R(R^2 - 1)}$$

where $\sigma^i_j$ is the rank of $y^i_j$ in the set of annotations $\{y^i_j\}$. By evaluating the correlation between each pair of annotators, we can compute the average agreement for individual annotators. The agreement can also be used to distinguish between the different types of annotators described in Section IV. Adversarial annotators will tend to have negative agreement with the good annotators, whereas spammers (random annotators) will tend to have an average agreement near to 0.

The mean inter-annotator agreement was obtained for each annotator by computing their average Spearman correlation with all other annotators assigned the same task. From Figure 7a it can be seen that participants carrying out the classification task had significantly lower agreement than those from the ranking and scoring tasks. There were also relatively few negatively correlated annotators, with the large majority obtaining positive average correlations.

Figure 7b is the result of bootstrapping on the set of mean inter-annotator agreements from each task. It shows 1000 bootstrapped sample estimates of the mean and standard deviation for each task. The separation of the classification task from the other two is clear, with the annotators showing only small variations in agreement. The scoring and ranking tasks on the other hand show more variation in addition to a higher mean agreement.

As the results show, the ranking task obtained the highest level of agreement among the annotators. The participants from the classification task obtained significantly lower agreement. This is partly to be expected due to the nature of the task, as there is no notion of the degree to which classifiers agree on an instance, only whether they agree or disagree.

C. Reliability of Annotators

In order to test the reliability of the annotators, we calculated their accuracy in relation to the consensus scores on both the original and rotated images. Analysis of variance was then used to compare the consensus accuracy on the original and rotated images. This gives an idea of how consistently the annotators maintained their accuracy throughout the duration of the tasks. The results are shown in Table II.

None of the tasks showed a statistically significant difference between the accuracies on the original and rotated images. The ranking task showed a particularly strong similarity between the two sets of accuracies, indicating that the annotators
were reliable in estimating the degree of clumpiness present in the images.

An indication of the correspondence between annotators from each of the tasks is found by considering the plot in Figure 8. The plot shows the combined scores from both the original and rotated images, sorted in increasing order for each task and translated to a common scale. It can be seen that the scores from each task encompass a similar range, starting from above 0 and ending near to 1. Each task resulted in a similar proportion of scores at the higher end of the scale, with more divergence between the tasks around the middle. This suggests that the images varied in difficulty, from those which were quite obviously clumpy/not-clumpy to the more ambiguous images around the middle range, where the annotators can assign both high and low scores. Very few annotators assigned scores of 0 to an image and the plot indicates that at least some degree of clumpiness was distinguished in each image.

<table>
<thead>
<tr>
<th>TASK</th>
<th>$F(1,62)$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLASSIFICATION</td>
<td>0.554</td>
<td>0.458</td>
</tr>
<tr>
<td>SCORING</td>
<td>0.36</td>
<td>0.85</td>
</tr>
<tr>
<td>RANKING</td>
<td>0.003</td>
<td>0.953</td>
</tr>
</tbody>
</table>
VII. OPTIMISING WEIGHTING PARAMETERS FOR MAJORITY VOTING

An important consideration when estimating the consensus annotations is how to weight the annotators according to their quality. The results reported thus far have given equal weight to annotators, but it is clear from the ROC curves shown in Figure 5 that both spammers and adversarial annotators are present. Using the majority vote accuracy on the gold standard as the objective, we optimised weighting parameters for the annotators using an evolutionary algorithm. The majority vote estimate of an instance was obtained using

\[ y_i = \sum_{j=1}^{R} w_j y_{ij} \]  

(11)

where \( w_j \) is the weighting parameter for annotator \( j \). The parameters themselves were drawn from the interval \([-1, 1]\) and constrained so that \( \sum_{j=1}^{R} |w_j| = 1 \). This normalises the majority vote estimates and allows adversarial and spamming annotators to be assigned negative and zero weights respectively. At each generation the algorithm randomly perturbed the annotator weights of selected members of a population of weights, retaining the best members of the population for the succeeding generation.

Figure 9 is a plot of the optimised annotator weightings for each of the tasks. It can be seen that annotators were assigned a range of weights, indicating differences in quality. Annotators assigned weights close to 0 can be assumed to be spammers, while those assigned negative weights are more adversarial. It is interesting to observe that in each task just a few annotators are assigned significantly larger weights than the majority, but each task had a few effectively adversarial annotators. Also, the classification task had the most spammers. By comparing the mean inter-annotator agreements, it was found that the annotators assigned positive weights also tended to be more strongly correlated than those with non-positive weights. This was true for all three of the tasks, with the increase in mean inter-annotator agreement between the non-positive and positive annotators being 0.02, 0.17 and 0.23 for the classification, scoring and ranking tasks, respectively. These results suggest that optimising the accuracy on the gold standard also indirectly optimised the inter-annotator agreement.

The majority vote estimates from all three of the tasks were able to obtain perfect accuracy on the gold standard.
optimised weighting parameters. For comparison, the standard (equally-weighted) consensus accuracy was 0.86, 0.57 and 0.86 for the classification, scoring and ranking tasks, respectively. Given that the annotators tended to be reliable in maintaining their accuracy, this means a significant improvement in the overall quality of the consensus estimates can be expected.

VIII. CONCLUSION

The use of citizen science for microscopy image annotation was shown to be viable. The annotations can be used to reliably characterise the distribution of clumpiness within the images. They could also provide additional insights by indicating images with highly variable annotations, which may be due to anomalies within the cell. The results also demonstrate that it is possible for a relatively large number of image annotations to be obtained from a comparatively small number of non-expert annotators.

Although the tasks required significant effort from the participants due to the variation and complexity of the images, the accuracy of the annotators was still generally high. The annotator estimates on the gold standard compared favourably with the expert annotations overall. Annotators from each task were also shown to be reliable in their estimates, with those from the ranking task in particular showing a strong similarity between the original and rotated image annotations.

A significant improvement in accuracy can be obtained by optimising annotator weighting parameters. This was demonstrated using an evolutionary algorithm to improve the accuracy of the consensus estimates on the gold standard. Using the optimised parameters, the majority votes for each task obtained perfect accuracy on the gold standard. It was also found that annotators with higher mean inter-annotator agreement tended to be assigned greater weight, suggesting a correlation between the accuracy on the gold standard and the overall agreement among the annotators.

The annotations obtained from the experiment demonstrate that the type of task presented to annotators had a significant impact on the quality of the resulting data. All three of the tasks showed clear differences in accuracy and inter-annotator agreement, with the ranking task obtaining the best overall performance.

REFERENCES