User data aggregation for StallCatchers.com Crowdsourcing Initiative

StallCatchers Team:
Mohammad Haft-Javaherian (mh973), Silvia Zhang (yz937), Fei Xia (fx43).

1 Introduction

The goal of crowdsourcing and machine learning is to combine the human knowledge and machine computing to aid research and problem solving in areas where combined efforts of human and machine are greatly needed. Crowdsourcing creates new training data for machine learning. For instance, Zhu et al. [1] used Amazon Mechanical Turk [2] to acquire ground truth for their generative adversarial networks. However, crowdsourcing faces a general problem that expertise of the majority could be questionable. When the crowd makes contradictory predictions or even sometimes crowd is wrong [3], it is a challenge for scientists to determine the optimal result. Therefore, it is necessary to develop an algorithm to adaptively combine crowd results and make a decision based on individuals precisions from previously known training data.

2 Dataset

Data were acquired through StallCatchers.com, which is a crowdsourcing website co-developed by several research institutes to study correlation of naturally stalled vessels in Alzheimer's disease. StallCatchers.com divided the recorded data into a calibration set and a validation set. Calibration set contains vessels studied by an expert who gave a definite answer with descriptions stating why she believes in the selected correct answer. The usage of this dataset includes education, user training, and crowdsourcing calibration process. Each user playing StallCatchers will see one of these calibration vessels after each 5 to 25 unknown vessels, based on the user experience level, in order to capture the user accuracy for further analyses. In addition, a validation set was created for the purpose of crowdsourcing methodology validation. It validates the complete automated analysis of the *in vivo* multiphoton microscopy images of mouse brain vasculature in wildtype and Alzheimer's disease mouse model (APP PS1/Swd), in which StallCatchers is the last segment of that pipeline.

2.1 Data acquisition

StallCatchers.com has more than 6000 users and a hundred thousand of vessels. Players start with a training video to learn how to play the game followed by a hands-on session that after each round of the game, they receive details feedback describing the correct answer and the location of stalled red blood cells. Each player at each round of the game sees a 3D stack image of brain blood vessels, which is acquired using multiphoton microscopy. The targeted vessel was outlined for users to determine whether the vessel was stalled or flowing. Due to the rastering mechanism of the multiphoton microscopy, different slices were acquired within half a second. This fact allows users to observe the movement of red blood cells as black patches and help them to discretize between stalled and flowing vessels. In order to record user accuracy, after several unknown vessels, they will be tested on one of the calibration vessels with known answers.

2.2 Dataset structure

The StallCatchers data is managed using MySQL. The users data can be exported as comma separated value (csv) text files that each line of record contain timestamp, vessel ID, user ID, user answer (0 = flowing vessel, 1 = stalled vessel), and Calibration flag (0 = novel vessel and 1 = calibration vessel).

3 Related Work

In the past, there have been several major approaches in machine learning to process and improve crowdsourcing data. One adaptive method is called active learning, which selects a subset of data with unknown labels for users to label to improve the accuracy of the overall algorithm. Beygelzimer et al. [4] described how they evaluated "values" from unlabeled points, which are the most crucial to determine the labels for the rest of unlabeled samples. Then, by giving users only the samples that were weighted the most, they were able to acquire extra training points that defines the optimal algorithm. On the other hand, with users giving us controversial decisions on some of the samples, it is necessary for us to evaluate the qualities of different users. Quality estimation, which is another common method in analyzing crowdsourcing data, has several approaches in machine learning to achieve this goal. In estimating user quality, expectation maximization, or EM has been commonly adopted in estimating maximum likelihood of unknown parameters in a given model. In a couple papers, EM was used in maximizing user error rate [5] to determine which user to eliminate from analysis, or in maximizing labeler accuracy in correlation to image difficulty [6] to determine which sample will be given to which user to achieve maximal accuracy overall. There has also been an approach that utilizes EM to iterate between defining a gold standard and measuring user accuracies on that gold standard [7], which ideally gives a guideline for user expertise.

4 Data pre-prossessing

Data were acquired through an online game stallcatchers.com, which has more than 6,000 users. Users first get a website-based hands-on training before labeling vessels to record their accuracies. Then they will be given unlabeled vessels to label. After several unlabeled vessels, they will be tested on one of the calibration vessels with known answer. We first loaded the website user activity log using MySQL into a CSV file. It is essentially a $K \times 5$ matrix where each row represents one of K time-stamps recorded, and each column represents timestamp information, vessel ID, user ID, user answer (0 = flowing vessel, 1 = stalled vessel), and calibration flag (0 = novel vessel, 1 = calibration vessel). We then converted the times-stamp strings into number of seconds. After this process, we converted the $K \times 5$ matrix into a $L \times P$ matrix where L rows represent L vessels and columns represent the opinions from P users. For opinion such as "Not available", it is treated as "NaN" in the matrix. As not all of the users submitted opinions for all the vessels, the matrix we have as the input is a very sparse matrix. So the data I/O would be:

Data input: an $N \times P$ matrix with rows representing the vessels and columns representing the user opinions

Data output: an $M \times 2$ matrix where M is the total number of vessels we need to determine whether stalled or not, column 1 has information of vessel labels and column 2 has the information of the probability estimation of being stalled.

5 Method description

In this section, we will describe the baseline method in addition to the proposed method and its variations that we developed to improve its accuracy. Then, we will discuss a third group of methods that we explored and their results at this stage. We implemented our algorithms in MATLAB 2017a.

5.1 Baseline algorithm

Due the fact that the probability of stalled vessel incidence is less than 0.5% for wild type mice and 2% for the current Alzheimer's mouse model, a reasonable number of false positives is manageable using second round of the lab expert quality control over positive detections. On the other hand, false negatives are not tractable because quality control will not apply to them due to extremely large number of negative results. The baseline approach will be the current method that is adopted for aggregating the user data. The answers of users for each vessel are weight-averaged using user accuracy [7] metric called d' (calculated by $d' = Z_{hit\ rate} - Z_{false\ alarm\ rate}$, where Z indicates the inverse of cumulative distributed function of Gaussian distribution). The weighted average was then thresholded using threshold level, e.g. 50%, which was found using cross validation over the data acquired during the alpha test of the game in 2016. On the other hand, ROC curve based on different threshold levels can give more information and freedom. Note that the d' should be calculated based on the known answer, e.g. website training data, and will be applied to the validation data as a known parameter.

5.2 The first proposed algorithm (STAPLE)

Assume we have R raters in total. Let $p = (p_1, p_2, ..., p_R)^T$ be a column vector of these R raters, describing the sensitivity, and let $q = (q_1, q_2, ..., q_R)^T$ be a column vector for specificity. Both p and q are used to characterize the performance of raters. Let D be an $S \times R$ matrix for the binary decisions made by each rater for each one of the S vessels. Let T be an indicator vector to represent the true labels, which are unknown in many crowdsourcing cases. Therefore, the probability mass function of the complete dataset can be $f(D, T|p, q)$. Here we want to maximize the complete data log likelihood function with estimation of p, q, using the M-step of the EM algorithm [4]. The M step has the following optimization problem by the Eq. [1].

$$\left(\mathbf{p}^{(k)}, \mathbf{q}^{(k)}\right) = \arg\max_{\mathbf{p},\mathbf{q}} \sum_j \sum_i E\,\Big[\ln f(D_{ij} \,|\, T_i, p_j, q_j) \,\Big|\, \mathbf{D}, \mathbf{p}^{(k-1)}, \mathbf{q}^{(k-1)} \Big] \tag{1}$$

The performance parameters are updated by the Eq. [2] and Eq. [3]. W is the conditional probability of the stalled vessels (also referred as weight variable in EM algorithm).

$$p_j^{(k)} = \frac{\sum\limits_{i:D_{ij}=1} W_i^{(k-1)}}{\sum\limits_i W_i^{(k-1)}}, \tag{2}$$

$$q_j^{(k)} = \frac{\sum\limits_{i:D_{ij}=0} \left(1 - W_i^{(k-1)}\right)}{\sum\limits_i \left(1 - W_i^{(k-1)}\right)} \tag{3}$$

For the E step, we estimate the conditional probability of the stalled vessels, W, given the rate decisions and previous performance parameters for each vessel i. At each iteration k, we have

$$W_i^{(k-1)} = \frac{a_i^{(k-1)}}{a_i^{(k-1)} + b_i^{(k-1)}} \tag{4}$$

where,

$$a_i^{(k)} = f(T_i = 1) \prod_{j:D_{ij}=1} p_j^k \prod_{j:D_{ij}=0} \left(1 - p_k^{(k)}\right) \tag{5}$$

$$b_i^{(k)} = f(T_i = 0) \prod_{j:D_{ij}=1} q_j^k \prod_{j:D_{ij}=0} \left(1 - q_k^{(k)}\right) \tag{6}$$

Note that the $f(T_i)$ is the prior probability of T_i and is treated as a binary variable, so $f(T_i = 1) = 1 - f(T_i = 0)$. The E-step and M-step are iteratively until W converges.

5.3 Adaptation to StallCatcher dataset

5.3.1 Handling sparse dataset

In the original STAPLE algorithm, dataset was assumed to have no missing entry. However, this was not the case for us since each user was only given a limited number of vessels for analyses. Due to a large number of vessels needed to be labeled, each vessel was only given to a few users. In order to account for the missing entries in the input data, we omitted the users who have not participated in any of the vessels we are analyzing here and omitted the vessels that did not have a single label from the users.

5.3.2 Problem with more users than vessels

STAPLE algorithm made an assumption on having higher number of data points than number of labelers, which was most appropriate when handling image data. However, with crowdsourcing, number of users can grow to a much larger number. In the case of applying STAPLE algorithm, this causes calculation of conditional probability in Eq. [5] and Eq. [6] go to infinitely small numbers (i.e. $< 10^{-500}$), which leads to computer running out of bits and outputting incorrect values. To address this problem, we replaced all the values smaller than 10^{-500} with 10^{-200}, keeping the nature of infinitely small values without outputting faulty results.

5.3.3 Two different implementations of STAPLE algorithm

We attempted two different implementations of STAPLE algorithm to test their performances. In "STAPLE 1" method, we trained the program with our training dataset. Then we inputted the calculated sensitivity and specificity of each user to STAPLE algorithm on the validation dataset. In "STAPLE 2" method, we assumed no prior knowledge on the users performance and ran the algorithm directly on validation dataset. The differences in performance between the two are further addressed in the Results section.

5.4 Third group of methods

In conjugate with the baseline and the main idea of the STAPLE algorithm, we applied other accuracy metrics and their combinations in order to improve the overall performance of the aggregation algorithm. We tested the following metrics including Dice Index(DI), Jaccard Index (JI), sensitivity, specificity, or the combinations of sensitivity and specificity for weighting the user opinions. JI results the highest AUC.

$$sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$specificity = \frac{TN}{TN + FP} \tag{8}$$

$$precision = \frac{TP}{TP + FP} \tag{9}$$

$$Jaccard\ Index\ (JI) = \frac{TP}{TP + FP + FN} \tag{10}$$

$$Dice\ Index\ (DI) = \frac{2 \times JI}{1 + JI} \tag{11}$$

where, TP,TN, FP, FN refer to true positive, true negative, false positive,and false negative, respectively.

5.5 Improved algorithm with user history dependency

When collaborating with a psychologist in analyzing the user input from StallCatchers, we were informed that human accuracy could potentially be time dependent. Previously, we were running our algorithms assuming no time dependency and used all previous history to calculate user accuracy. To incorporate this new variable, we prepared a lookup table that includes each user's full history at any given time point. Then we calculated the Jaccard Index of each user when only considering the past certain number of clicks. This new accuracy metric for each user was then utilized to calculate probability of each vessel being stalled.

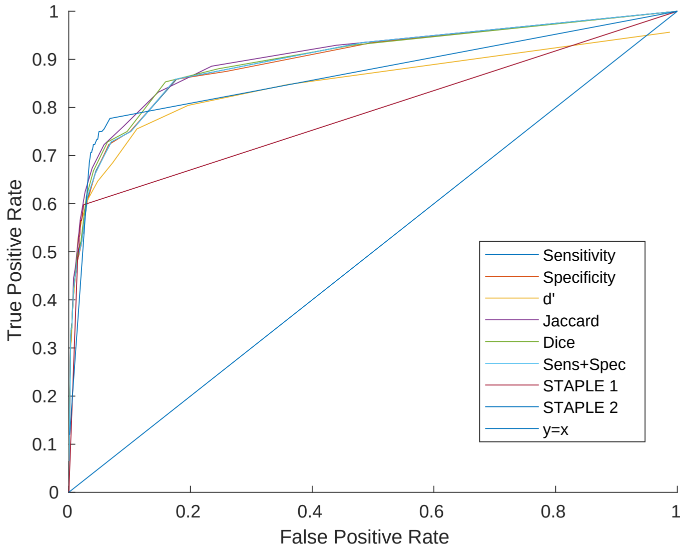

Figure 1: The receiver operating characteristic curve (ROC) curve of the different proposed method in addition to the baseline method applied to the validation dataset. Such result suggests that any of Dice Index, Jaccard Index, sensitivity, specificity, or the combinations of sensitivity and specificity could be used for crowd result aggregation. Note that the ROC values for those methods are quite similar and overlapping.

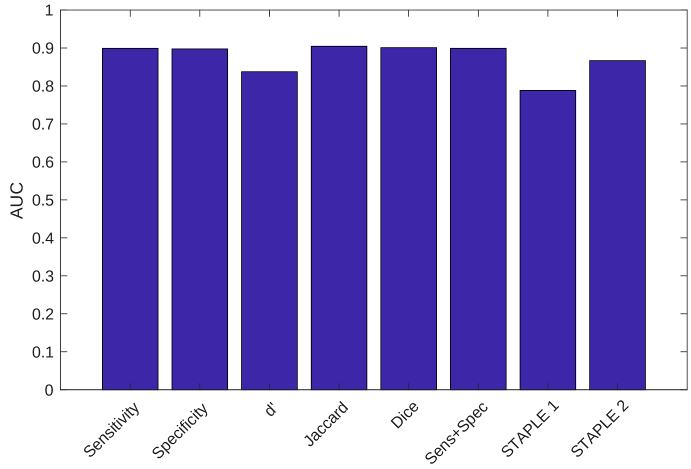

Figure 2: The the area under the ROC curve (AUC) for the different proposed method in addition to the baseline method applied to the validation dataset. Overall, even though the STAPLE 2 method outperforms STAPLE 1 method and the baseline method, STAPLE 1 was the worst method. Other methods are quite similar in terms of AUC values.

6 Results

The ROC curve in addition to the AUC bar plots based on our validation dataset are shown in Figure 1 and Figure 2, respectively. The results show that the baseline method delivers a reasonable result even better than the proposed method. On the other hand, the third group of methods demonstrated the best results, specially JI. Their accuracies are quite similar as shown in Figure 2. Up till here, we believe JI and the sens+spec weighted method, which use sensitivity to weight the positive signals and specificity to weight the negative signal, are the best options.

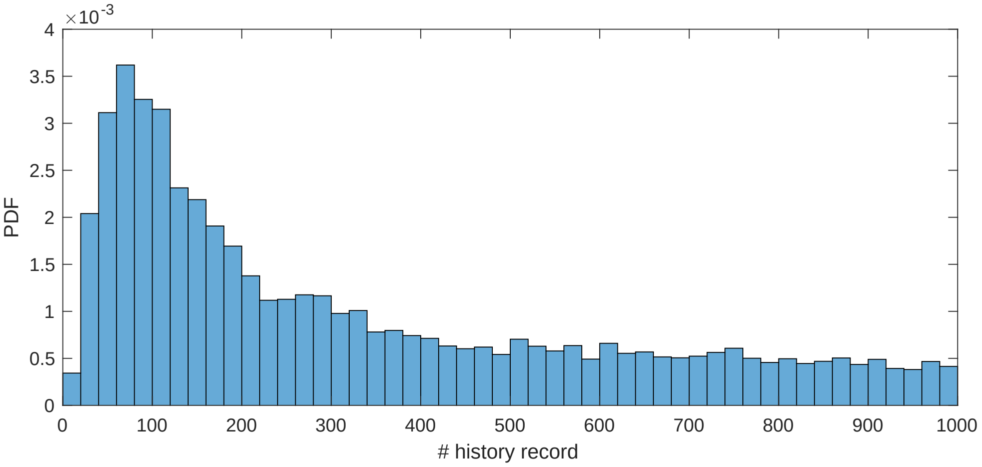

Figure 3: Number of available calibration records prior to current time point. PDF: probability density function.

Interestingly, the STAPLE 2, which does not assume any prior knowledge about the users' sensitivity and specificity performed much better than the STAPLE 1, which used the training dataset as a source to obtain the prior knowledge about the user accuracies. This result may imply that the users' accuracies depend on the nature of multi-photon microscopy *in vivo* images, particularly the image quality. The results based on the STAPLE 1, method, are closest to the total chance line (y=x, which imply the equal 50% chance i.i.d. uniform chance for randomly classifying each vessel). Therefore, these metrics need to be calculated for each project individually. This result also suggests the possibility of these metrics being time-dependent, which means that user accuracy changes over time.

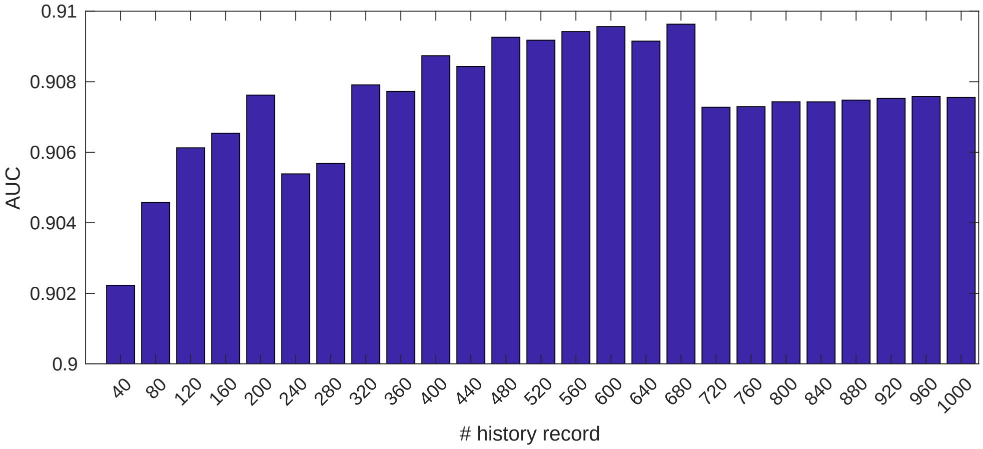

Figure 4: Number of records considered for user accuracy measurement based on Jaccard Index.

To incorporate the time-dependency of user accuracy, we implemented the last algorithm we proposed in the method section. Figure 3 shows the probability density function of percentage of users reaching different number of records. From the figure, it is straightforward to see that majority of the users have a history of 20 ∼ 180 records whereas there are certain number of consistent “super” users that contributed from 300 to 1000 vessels, and even some with 40’000 vessel records. The AUC calculated with different thresholds of latest user inputs were shown in Figure 4. From the AUCs, we can see that the overall AUC from the past history were relatively consistent in the range between 90% to 91%. However, the AUC did improve gradually when we considered more than 40 of past recent user records. Surprisingly, the AUC went lower when the past user records reached over 200, where we thought differentiated a super user from normal users. The decrease from 680 to 720 might indicate that even for those super users, they had lower accuracy at the beginning but gradually improved over the course of this study.

7 Limits and Future plans

There are several limitations in our methods proposed so far. First of all, we have a comparatively low number of vessels in the training dataset compare to that of the test dataset. Secondly, the difference between the stalled vessel distribution within the test is very different from that of in the train dataset ($\sim 2\%$ *vs.* $\sim 25\%$). This indicates that the players may be influenced by this as a prior knowledge which could potentially affect their performance measurement. Thirdly, as super players introduce new behaviors such as memorization that causes 100% user accuracy, which saturates baseline threshold method. We have to do a second round of opinion collection in order to have a complete dataset. Finally, some occasional players can introduce noisy user accuracy results.

To address the above limitations, our future plans will be: first, we can study the minimal number of the required data points per vessel; second, we can fix the parameters as a function of task difficulty. We can also train the model with larger test and training dataset; third, we can extract the model confidence to be used for post-processing in the lab validation; and also we can develop smart load distributions to users in order to assign hard-mode tasks to super users.

8 Conclusions

StallCatchers is a promising crowdsourcing approach to solve the stalled vs. flowing classification problem. To gather the opinions from StallCatchers users, various methods are available for user data aggregations and analysis. With all our explorations with different variations of STAPLE method, sensitivity, specificity, precision, Jaccard Index, Diced Index and user-history considerations, the method based on Jaccard Index with 700 user's history record provides the most promising approach. We expect more future methods to be developed in order to use human computation recourses wisely.

References

[1] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." arXiv preprint arXiv:1703.10593 (2017).

[2] Amazon Mechanical Turk - Welcome. [Online]. Available: https://www.mturk.com. [Accessed: 19-Sep-2017].

[3] Prelec, Draen, H. Sebastian Seung, and John McCoy. "A solution to the single-question crowd wisdom problem." Nature 541.7638 (2017): 532-535.

[4] Beygelzimer, Alina, Sanjoy Dasgupta, and John Langford. "Importance weighted active learning." Proceedings of the 26th annual international conference on machine learning. ACM, 2009.

[5] Joglekar, Manas, Hector Garcia-Molina, and Aditya Parameswaran. "Comprehensive and reliable crowd assessment algorithms." Data Engineering (ICDE), 2015 IEEE 31st International Conference on. IEEE, 2015.

[6] Warfield, Simon K., Kelly H. Zou, and William M. Wells. "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation." IEEE transactions on medical imaging 23.7 (2004): 903-921.

[7] V. C. Raykar et al., Supervised learning from multiple experts, in Proceedings of the 26th Annual International Conference on Machine Learning - ICML 09, 2009.

[8] Whitehill, Jacob, et al. "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise." Advances in neural information processing systems. 2009.