Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
An Adjunct to the Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing

Behavior-Based Quality Assurance in Crowdsourcing Markets

Michael Feldman, Abraham Bernstein
University of Zurich, Department of Informatics
{feldman, bernstein}@ifi.uzh.ch

Abstract

Quality assurance in crowdsourcing markets has appeared to be an acute problem over the last years. We propose a quality control method inspired by Statistical Process Control (SPC), commonly used to control output quality in production processes and characterized by relying on time-series data.

Behavioral traces of users may play a key role in evaluating the performance of work done on crowdsourcing platforms. Therefore, in our experiment we explore fifteen behavioral traces for their ability to recognize the drop in work quality. Preliminary results indicate that our method has a high potential for real-time detection and signaling a drop in work quality.

Introduction

To date, most common quality control policies in crowdsourcing focus on controlling errors via result aggregation following the elicitation of multiple answers for the tasks. Another common mechanism is based on ground truth embedding into the workflow, such that a worker's creditability can be estimated based on her responses (Oleson et al. 2011). Our study is adjacent to the one proposed by Rzeszotarski and Kittur (2011). This study proposes a technique that captures per task aggregated behavioral traces of crowd-workers (e.g., mouse movements, scrolling) and by training machine-learning models to predict their performance in future tasks.

We address a commonly used crowd-work setup where the HIT (Human Intelligence Task) consists of sequential and identical subtasks such as item recognition or content labeling. This study considers the HIT as a continuous process, where crowd-workers, as a result of training, achieve performance improvement, with a subsequent productivity period. Eventually, their performance drops as a result of fatigue, distraction or any unanticipated causes. Therefore, we propose a quality assurance technique that is not limited to inferring the performance quality based on aggregated behavioral traces of previous tasks but rather on continuous-time analysis of the behavior of the crowd-workers.

Statistical Process Control is a well-known statistical field for quality control and improvement in manufacturing and business processes. Statistical control in this context refers to a stabilized process in which only common causes of variation remain and all special (unanticipated) causes of variation have been removed. An XmR chart is an individual's control chart that is used for tracking the time-series of a process in order to determine whether that process is in statistical control and may be considered as stable. When a process is considered stable, it experiences only common-cause variability. When a process is not in control, special-cause conditions can be causing non-stability (Montgoemery 2007). Applied to time-series data, the method defines limits of acceptable output while an extension over the limits, or systematic patterns within, indicate a new source of variation. Therefore, we calculate a metric for each behavior trace (e.g., number of mouse clicks per time) and look for patterns that will indicate a drop in performance. This allows us to recognize patterns in behavioral data that are likely to signal deterioration in work quality. Note, as this work is research in progress we present the preliminary results of our analysis; a full analysis will be presented in future work.

The quality assurance method proposed in this study has the following potential impact: First, while assuming a sequence of micro tasks, the method is applicable to single, general, and complex tasks. Second, requesters get the opportunity to control the work quality almost in real-time due to continuous behavioral traces. Lastly, the method will potentially allow for a semi-automatic redesign of the task workflow, such that breaks can be suggested or results of questionable quality are doubled by additional crowd-worker.

Our approach faces some challenges. For instance, we encountered a technical problem, where streaming data can slow down the users' device. Another challenge is preventing gaming the system. However, building the system such that it considers various behavior traces as well as dynamically learning the behavior of individuals will mitigate this risk.

Method

A total of one hundred fifty participants took part in the study. The participants were randomly drawn from the Mechanical Turk platform's pool of workers and were paid 3 USD for an experiment lasting about twenty minutes. Most participants originated from the US (59%), India (15%), China (12%), and the rest are primarily from European countries such as Germany, France and Russian Federation.

The study consists of five tasks; each is a sequence of sub-tasks that are very similar to each other and resemble tasks often to be found in crowdsourcing labor platforms. The designed tasks included restoration of distorted text, recognition of specified details in a picture, content filtering and constrained ordering, image recombination, and content tagging. While performing these tasks, fifteen behavioral traces such as mouse speed and acceleration, duration of mouse clicking, scrolling speed, typing speed, were tracked. The performance for each sub-task was recorded in order to recognize patterns in the participant's behavior that correspond with the drop in performance.

Preliminary Data Analysis

The analysis was conducted using a separate XmR chart for every participant. Preliminary results show that patterns in a participant's behavior may indicate a significant shift in work quality (see Figure 1). Additionally, it seems that a substantial increase in variance better predicts the potential for a decrease in quality than behavioral performance per se. To illustrate this, we present the performance and behavior metric of a particularly interesting individual, and discuss the main findings.

The top of Figure 1 presents the XmR chart with behavioral metric of mouse direction changes during continuous mouse movement. On the bottom, the participant's task performance is presented, where each horizontal line represents the performance for specific subtask. The drop in quality is reflected in the behavioral trace, since multiple data points extend beyond the upper control limit of the XmR chart. Moreover, the variance of the behavioral metric increases over time and reaches the highest level towards the end of the process. Therefore, it is reasonable to assume that the drop in performance is related to both (i) an increase in variance and (ii) exceeding the control limits. We did not perform a statistical test yet but these phenomena appear to exist for different individuals who were tracked using the aforementioned behavioral metric.

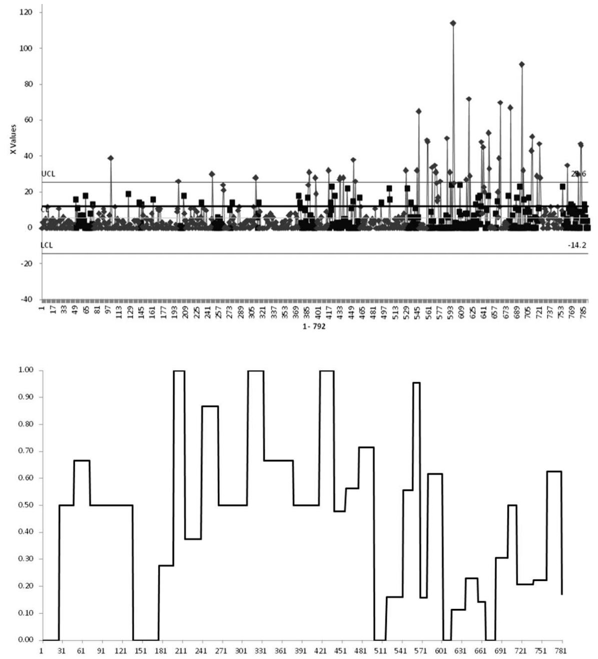

Figure 1: Quality control chart of a behavioral metric (on the top), versus task performance over time (on the bottom).

Conclusion and Future Work

This study proposes a method for quality control that tracks different behavioral traces of crowd-workers and analyzes them by means of Statistical Process Control to identify the significant changes in performance over time. Our preliminary analysis for one behavioral metric demonstrates tendencies based on the data of a single participant. For future work, we will extend the data analysis to identify significantly similar patterns between different individuals for each of the tracked behavioral metrics. Additionally, we will explore multiple metrics simultaneously to determine whether more complex patterns, based on different behavioral traces, can be found. We are confident, that the results of this analysis will pave the way for behavior-driven quality assurance tools that are able to track the performance of individuals and to redesign the workflow automatically if necessary.

References

Oleson, D., Sorokin, A., Laughlin, G. P., Hester, V., Le, J., and Biewald, L. 2011. Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. *Human computation, 11, 11.*

Montgomery, D. C. 2007. *Introduction to Statistical Quality Control*. John Wiley & Sons.

Rzeszotarski, J. M., and Kittur, A. 2011. Instrumenting the Crowd: Using implicit Behavioral Measures to Predict Task Performance. *In Proceedings of the 24th annual ACM symposium on User interface software and technology,* 13-22. ACM.