Crowdsourcing Control: Moving Beyond Multiple Choice

Christopher H. Lin  Mausam  Daniel S. Weld
Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195
{chrislin,mausam,weld}@cs.washington.edu

Abstract

To ensure quality results from crowdsourced tasks, requesters often aggregate worker responses and use one of a plethora of strategies to infer the correct answer from the set of noisy responses. However, all current models assume prior knowledge of all possible outcomes of the task. While not an unreasonable assumption for tasks that can be posited as multiple-choice questions (e.g. \(n\)-ary classification), we observe that many tasks do not naturally fit this paradigm, but instead demand a free-response formulation where the outcome space is of infinite size (e.g. audio transcription). We model such tasks with a novel probabilistic graphical model, and design and implement LazySusan, a decision-theoretic controller that dynamically requests responses as necessary in order to infer answers to these tasks. We also design an EM algorithm to jointly learn the parameters of our model while inferring the correct answers to multiple tasks at a time. Live experiments on Amazon Mechanical Turk demonstrate the superiority of LazySusan at solving SAT Math questions, eliminating 83.2% of the error and achieving greater net utility compared to the state-of-the-art strategy, majority-voting. We also show in live experiments that our EM algorithm outperforms majority-voting on a visualization task that we design.

1 Introduction

Crowdsourcing marketplaces (e.g., Amazon Mechanical Turk) continue to rise in popularity. Hundreds of thousands of workers produce a steady stream of output for a wide range of jobs, such as product categorization, audio-video transcription and interlingual translation. Unfortunately, these workers also come with hugely varied skill sets and motivation levels. Ensuring high quality results is a serious challenge for all requesters.

Researchers have studied quality control extensively for the case of simple binary choice (or multiple choice) questions. A common practice is to ask multiple workers and aggregate responses by a majority vote [Snow et al., 2008]. Several extensions have been proposed that track the ability of individual workers while estimating the inherent difficulty of questions [Dai et al., 2010, Whitehill et al., 2009]. These methods typically outperform majority vote and achieve a much higher accuracy.

A key drawback of prior decision-theoretic approaches to quality control is the restriction to multiple choice questions, i.e., jobs where every alternative answer is known in advance and the worker has to simply select one. While many tasks can be formulated in a multiple-choice fashion (e.g. \(n\)-ary classification), there a large number of tasks with an unbounded number of possible answers. A common example is completing a database with workers’ help, e.g., asking questions such as “Find the mobile phone number of Acme Corporation’s CEO.” Since the space of possible number of answers is huge (possibly infinite), the task interface cannot explicitly enumerate them for the worker. We call these tasks open questions.

Unfortunately, adapting multiple-choice models for open questions is not straightforward, because of the difficulty with reasoning about unknown answers. Requesters, therefore, must resort to using a majority-vote, a significant hindrance to achieving quality results from these more general open questions.

Our paper tackles this challenging problem of modeling tasks where workers are free to give any answer. As a first step, we restrict these tasks to those which have exactly one correct answer. We make the following contributions: