Samasource Worker Quality – Data Mining Project

Team Members

Chaya Nayak	**Chaya.J.Nayak@gmail.com**	**414.690.6980**
Laura Gerhardt	**Laura.Gerhardt@gmail.com**	

Background

Samasource, a non-profit organization operating in the San Francisco Bay Area, works to provide enterprise data solutions to organizations by connecting them to the untapped potential of women and youth in poverty in India and Africa. Working with clients such as Google, Trip Advisor, and Getty Images, Samasource provides workers that aid in the development of supervised machine learning algorithms in order to categorize data and improve organizational efficiency.

One important consideration for Samasource in the work that they do, is the quality of the final product they provide to their clients. Currently, Samasource hosts client projects on the Samahub, which breaks these projects into individual data points or micro-work. An agent on the Samasource team, works through sets of data throughout their workday and their individual quality is determined through Gold. Gold is a training set with correct answers that are fed into the agent's work stream. These Gold tasks are fed to agents at various points within their work:

1. **Onset of work:** At the onset of work, agents are fed **Gauntlets**, or a series of gold tasks that determine whether they are qualified to work on the project. If workers do not qualify, they must work on Gauntlet tasks until their quality improves.
2. **Mid-work:** Throughout the work process and after the initial qualification, agents are given **Gold Bursts**. These gold bursts give the works 10 gold tasks that assess the quality of their work. Samasource uses these gold tasks to determine the overall accuracy of the data set.

After the workers complete a set of work, the work set goes through two more quality assurance measures. The first quality assurance measure is a Team Lead Review. During this process, leaders of the agent teams review a sample batch of data on the hub in order to judge total accuracy, and improve overall data set accuracy. After the Team Lead Review, QA Leads preform Smart Review which looks at the entire merged data set, and uses sort features, and knowledge of common mistakes in order to determine and improve worker quality.

One realization however, within Samasource's quality measurement platform is the time consuming nature of quality assurance. For this reason, Samasource has been working with their engineering team to develop a semi-supervised machine learning algorithm that classifies workers based on performance during gold bursts, and "gates" Workers who do not fall within a pre-determined quality threshold. The gating process temporarily removes workers from the data set, and places them on a training set, feeding them gold rather than real tasks. Within this training set, workers have to meet the pre-determined quality threshold, as well as prove to management that they understand the mistakes they made in the dataset. The managers will then un-gate the workers.

Samasource currently has a gating algorithm, and has to date used fake worker data in order to test the gating algorithm. As the algorithm nears implementation, the engineering team needs support and guidance in regards to ensuring that the gating algorithm is functioning and will actually improve overall worker quality, while not substantially increasing an agents time per task.

Project Details

- We will initially develop decision trees and clusters to further classify the performance levels of Samasource workers and identify any surprising trends.
- Based off of this classification, we will then seek to improve the existing Samasource algorithm to respond to both underlying worker characteristics such as pre-program entry performance and current performance on the assessment questions (gold) to minimize supervisor-lead remedial gating.

Project Goals

The goal of this project is to develop an algorithm for gating workers who dip below a chosen quality threshold in order to improve accuracy of supervised machine learning projects at Samasource. The secondary goal of this project is to minimize gating in order to increase efficiency of project completion.

The project will explore further questions such as: How well does gold predict overall dataset accuracy when gating is used? Will categorical gating increase accuracy of quality prediction?

Assessment of Project Goals

- Our resulting classification algorithm will perform better on a testing set than the existing algorithms.
- The algorithm will be sustainable across multiple SamaSource Projects.

Deliverables

We hope to deliver an algorithm that improves the worker's accuracy while minimizing their need to engage in answering gold-level questions.

Work Breakdown

Initial Data Analysis and Cleaning	**1 week**
Develop Feature Models for semi-supervised classification	**1 week**
Assessment and Finalizations	**1 week**

Declaration

This project is being jointly pursued by Laura Gerhardt and Chaya Nayak from the Goldman School of Public Policy for Info 290: Introduction to Data Mining