ChimpsThrow: Active Learning with Crowdsourcing for Automatic Data Categorization

Hohyon Ryu

School of Information, University of Texas - Austin

Automatic categorization or classification has been one of the major applications of supervised machine learning [1]. The performance of supervised machine learning heavily relies on the quality of the training corpus. The present study closely investigates how to **improve training corpus utilizing active learning and crowdsourcing to maximize the performance of automatic data categorization**. Crowdsourcing is a good tool to incorporate human labor to machine learning algorithm, but its reliability is always questionable as we cannot control the quality of the workers. To minimize the impact of possible erroneous or abusive feedback from crowdworkers, active learning is used to balance computing power with human labor [2,3].

This study presents **ChimpsThrow**, a crowdsourcing-based active learning system for efficient automatic dataset categorization. This project is based on a real-world problem of categorizing **13653 datasets** of **Infochimps**(1). The 22.4\% of the datasets were categorized by hands, and a neural network classifier was implemented to automatically categorize the rest of 10585 documents. The data that the machine learning algorithm predicted with low confidence was crowdsourced at Amazon Mechanical Turk to get labeled by human. The feedback of the crowdsourcing is used in the training set, and the neural network trains and predicts again for all the documents. Doing this iteratively, the presented system improves overall quality of automatic categorization.

(1) http://www.infochimps.com/

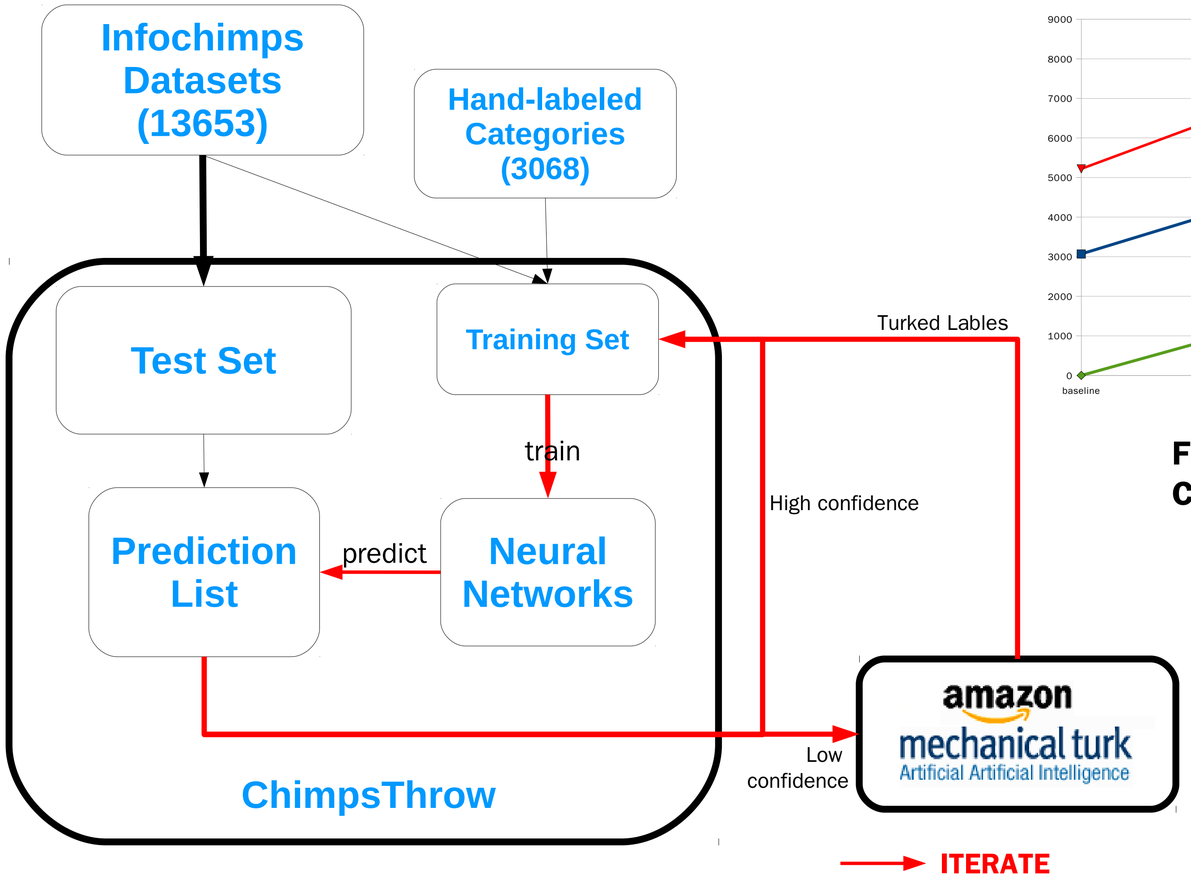

Figure 1. Overview of the Experiment

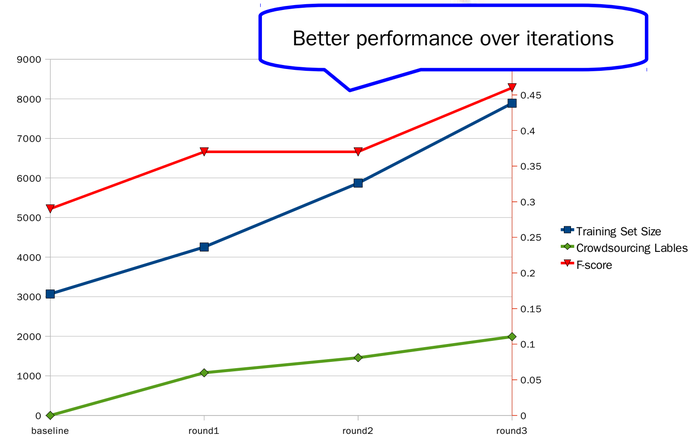

Figure 2. Training Data Size and Classification Performance

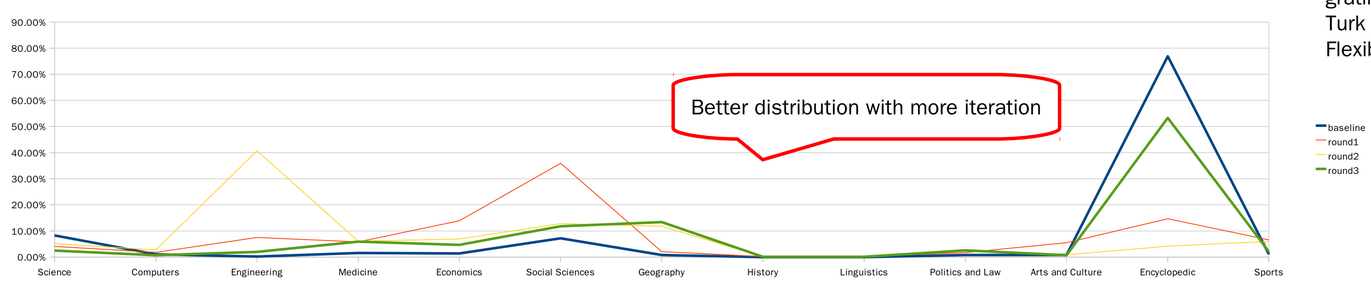

Figure 3. Classification Distribution per iteration

References

[1] F. Sebastiani, “Machine learning in automated text categorization,” ACM Computing Surveys, vol. 34, pp. 1–47, Mar. 2002.

[2] V. Ambati, S. Vogel, and J. Carbonell, “Active learning and crowd-sourcing for machine translation,” Language Resources and Evaluation (LREC), 2010.

[3] A. J. Quinn, B. B. Bederson, T. Yeh, and J. Lin, “CrowdFlow : Inte-grating Machine Learning with Mechanical Turk for Speed-Cost-Quality Flexibility”