2012

Dynamic Estimation of Rater Reliability in Subjective Tasks Using Multi-Armed Bandits

Alexey Tarasov
Dublin Institute of Technology, tarasovsaleksejs@gmail.com

Sarah Jane Delany
Dublin Institute of Technology

Brian Mac Namee
Dublin Institute of Technology

Dynamic Estimation of Rater Reliability in Subjective Tasks Using Multi-Armed Bandits

Alexey Tarasov, Sarah Jane Delany, Brian Mac Namee
School of Computing
Dublin Institute of Technology
Dublin, Ireland
Email: aleksejs.tarasovs@student.dit.ie

***Abstract*—Many application areas that use supervised machine learning make use of multiple raters to collect target ratings for training data. Usage of multiple raters, however, inevitably introduces the risk that a proportion of them will be unreliable. The presence of unreliable raters can prolong the rating process, make it more expensive and lead to inaccurate ratings. The dominant, "static" approach of solving this problem in state-of-the-art research is to estimate the rater reliability and to calculate the target ratings when all ratings have been gathered. However, doing it dynamically while raters rate training data can make the acquisition of ratings faster and cheaper compared to static techniques. We propose to cast the problem of the dynamic estimation of rater reliability as a multi-armed bandit problem. Experiments show that the usage of multi-armed bandits for this problem is worthwhile, providing that each rater can rate any asset when asked. The purpose of this paper is to outline the directions of future research in this area.**

***Keywords*-human computation; crowdsourcing; multi-armed bandits; learning from multiple sources; training data; supervised machine learning**

I. Motivation

The success of supervised machine learning depends heavily on the quality of training data. Getting target ratings can be especially challenging if correct ratings ("ground truth") are subjective, i.e. raters would tend to provide different answers due to a different level and area of expertise (for instance, in medical imaging [1]) or due to perceptual judgements (such as rating adult content [2] or emotion recognition from speech). Therefore, each training asset has to be rated by multiple raters, some of which can be noisy due to subjectivity of the task or a lack of care or attention to the task by the workers. In such setup the challenge of discovering which ratings are reliable arises [3]. The dominant, "static", approach to addressing this issue is first to collect all ratings and then to aggregate them (in order to get a single resulting consensus rating for each asset) as well as to estimate the rater reliability [4]. However, there are benefits to identifying good raters in the process of rating. Ratings will be of better quality and the consensus rating can be achieved faster, saving on cost in circumstances where rating involves monetary payment. This is considered "dynamic" estimation [5] but is not as common as the static approach.

II. Statement of problem

To estimate rater reliability we cast the rater selection problem as a *multi-armed bandit* (MAB) problem [6], which represents a task as a k-armed slot machine. Each arm on the slot machine can be pulled after which a numerical reward is received—the higher the reward, the better the arm. The task is to select arms to pull to maximise the reward. The rewards received by pulling each arm at each step of the algorithm are used to calculate the "quality" of each arm which inputs to the selection process. For our task, each available rater corresponds to an arm. At each iteration of the process we can choose rater(s) from the full rater population from whom to solicit ratings—asking a rater to provide a rating for an instance is equivalent to pulling an arm. We can set the reward received after selecting a rater (or pulling an arm) to be based on the accuracy of the rating received. Rater accuracy typically is unknown in scenarios where multiple raters are involved, but can be estimated, for example, using rater consensus.

III. Work to date

The comparison of different MABs in the task of dynamic estimating rater reliability was carried out. We performed a simulated rating experiment, where all ratings were collected in advance and instead of querying a rater in real time, we took an according value from the dataset. The results for rating jokes and movies [7], as well as rating emotional speech (to be published soon) present strong evidence that MABs are suitable for this task. MABs showed better performance than IEThresh [5], state-of-the-art algorithm for performing such estimation. Additionally, a rating tool [8] has been developed for rating in-house emotional speech corpus, which will be used in future experiments.

IV. Proposed work

Our work to date assumed that all raters are available to rate any asset at any time, however, it might not be true in real life. One of our next steps is to investigate how should

MABs be used when raters are sometimes unavailable to rate. If a rater is unavailable, other raters can be queried instead, the process might go on without that particular rating etc.

Another important problem is that MABs require performing a number of pulls before the quality of each arm gets estimated. It might lead to the pulling of inferior arms, i.e. asking noisy raters, who will provide inaccurate ratings. We plan to look into the question of how to correct ratings received at the early stage of the rating process. To correct ratings for "early" assets, additional ratings for them can be acquired from good raters, when their performance is estimated reliably.

Another set of experiments will be devoted to choosing number of raters to be asked at each step of the rating process. When the reliability is unknown, more raters should be queried. When good raters are discovered, only a small number of them might be sufficient.

These questions will be addressed before February 2013, when the start of the thesis write-up is planned. Expected submission—August 2013.

V. Motivation for participation

The main reason for the participation in the doctoral consortium is to get feedback from scientists, who need to rate a large amount of training data and are looking for the ways to cut time and cost associated with the rating process. We are also interested in the applicability of different social computing techniques to our problem.

Acknowledgements

This work was supported by the Science Foundation Ireland under Grant No. 09-RFP-CMS253.

References

[1] S. Cholleti, S. Goldman, A. Blum, D. Politte, and S. Don, "Veritas: Combining Expert Opinions without Labeled Data," Tech. Rep., 2008.

[2] P. Ipeirotis, F. Provost, and J. Wang, "Quality Management on Amazon Mechanical Turk," in *Procs of HCOMP*, 2010.

[3] E. Law and L. Von Ahn, *Human Computation*. Morgan&Claypool, 2011.

[4] V. Raykar, S. Yu, L. Zhao, G. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning From Crowds," *JMLR*, vol. 11, pp. 1297–1322, 2010.

[5] P. Donmez, J. Carbonell, and J. Schneider, "Efficiently Learning the Accuracy of Labeling Sources for Selective Sampling," in *Procs of KDD*, 2009.

[6] O. Maron and A. Moore, "Hoeffding Races: Accelerating Model Selection Search for Classification and Function Approximation," in *Procs of NIPS*, 1994, pp. 59–66.

[7] A. Tarasov, S. Delany, and B. Mac Namee, "Dynamic Estimation of Rater Reliability in Regression Tasks using Multi-Armed Bandit Techniques," in *Procs of MLHCC Workshop (ICML)*, 2012.

[8] J. Snel, A. Tarasov, C. Cullen, and S. Delany, "A crowdsourcing approach to labelling a mood induced speech corpora," in *4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (LREC)*, 2012, pp. 72–76.