2012-09-19

Dynamic Estimation of Rater Reliability in Regression Tasks using Multi-Armed Bandit Techniques

Alexey Tarasov
Technological University Dublin, tarasovsaleksejs@gmail.com

Sarah Jane Delany
Technological University Dublin, sarahjane.delany@tudublin.ie

Brian Mac Namee
Technological University Dublin, brian.macnamee@tudublin.ie

Recommended Citation

Tarasov, A , Delany, S. J. & Mac Namee, B. (2012) Dynamic Estimation of Rater Reliability in Regression Tasks using Multi-Armed Bandit Techniques, *Workshop on Machine Learning in Human Computation and Crowdsourcing,* in conjunction with ICML 2012.Edinburgh, Mar 26.

Dynamic Estimation of Rater Reliability in Regression Tasks using Multi-Armed Bandit Techniques

Alexey Tarasov ALEKSEJS.TARASOVS@STUDENT.DIT.IE
Sarah Jane Delany SARAHJANE.DELANY@DIT.IE
Brian Mac Namee BRIAN.MACNAMEE@DIT.IE
School of Computing, Dublin Institute of Technology, Kevin St., Dublin 8, Ireland

1. Introduction

The success of supervised machine learning depends heavily on the quality of training data. In some areas acquiring target ratings for training instances can pose a significant challenge due to the effort and high costs involved (Whitehill et al., 2009). One way of approaching the problem is to use crowdsourcing, which facilitates fast and inexpensive collection of high quality ratings (Ambati et al., 2010). However, the presence of unreliable raters can prolong the process, make it more expensive and lead to inaccurate ratings (Sheng & Provost, 2008). The dominant approach to addressing this issue is first to collect all ratings and then to estimate the ground truth ratings taking rater reliability into account (Raykar et al., 2010; Whitehill et al., 2009). A different approach, which estimates the ratings and reliability dynamically while raters rate, is considered by Donmez et al. (2009) who propose an algorithm called IEThresh. We consider this approach, the dynamic estimation of rater reliability, in this paper, as it can considerably mitigate the effect of rating noise and also keep costs to a minimum compared to the collection of all ratings.

To perform dynamic estimation of rater reliability we cast the rater selection problem as a *multi-armed bandit* (MAB) problem (Maron & Moore, 1994), 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 the best arms so as to accumulate the biggest 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 which 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 crowdsourcing scenarios but can be estimated, for example using rater consensus.

In this paper we show that MAB techniques are suitable for performing the task of the dynamic estimation of rater reliability. We focus on crowdsourcing scenarios where the ratings are numerical values typically in a scale, in contrast to other research in the area which concentrates on binary ratings (Brew et al., 2010; Donmez et al., 2009).

2. Methodology

We generated two datasets for our experiments. The first was extracted from the MovieLens 10M[1] corpus and consists of a subset of 278 movies rated by 20 raters in the range (1,5). The second, extracted from the Jester[2] corpus, contains a subset of ratings from 20 raters who have rated all 100 jokes in the corpus in the range of (-10, 10).

Caelen & Bontempi (2010) categorize MAB algorithms into four groups: (i) Gittins index policy MABs, (ii) probability matching MABs, (iii) semi-uniform MABs and (iv) upper-confidence bound MABs. For our evaluation we select representative algorithms from groups (iii) and (iv) as the most popular and successful MAB algorithms. These are the ϵ-first approach (Vermorel & Mohri, 2005) and KL-UCB (Garivier & Cappé, 2011) respectively. We include two baselines: an approach that randomly selects raters, referred to as *Random*, and an approach, referred to as *Overall-Best*, that always uses the subset of raters deemed by Raykar's algorithm (Raykar et al., 2010) to be the most reliable. We compare these approaches with IEThresh,

[1] http://www.grouplens.org/node/73
[2] http://goldberg.berkeley.edu/jester-data/

the current approach to dynamic estimation of rater reliability. It should be noted that IEThresh is not regarded as an MAB technique by its authors, however, it could be considered a upper-confidence bound MAB technique.

We conducted a simulated crowdsourced rating experiment as follows:

1. Selection of assets: Assets were presented for rating one by one; the order of presentation was selected randomly, with results averaged over 100 runs.

2. Selection of raters: Each MAB technique was used to select N raters to rate the selected asset. In most MAB techniques this involved selecting the N raters with the highest reliability score at that point in the process. The ϵ-first approach, however, uses random selection at certain points in its process to encourage exploration. In some cases, small amendments had to be made to the MAB algorithms to allow the selection of $N > 1$ raters at a time. IEThresh is designed not to select the top N raters but to use all raters who achieve a threshold reliability, which means that the performance of IEThresh is the same regardless of N.

3. Calculating the reward: The consensus rating for the asset is estimated as the average of the all ratings received for the asset. Each rater's reward is calculated as a normalised difference between the consensus rating and the rating provided by that rater. Thus, the closer a rating is to the consensus, the more reliable the rater is deemed to be.

4. Updating the rater reliabilities: The appropriate MAB strategy is used to update the reliability of the rater based on rewards received by him for all assets he has rated up to that point in the process. The process iterates until all assets are rated.

The performance of each strategy is measured as the average absolute difference between the consensus rating and the ground truth rating across all assets rated. With no actual ground truth existing, we used the approach presented by Raykar et al. (2010) to generate ground truth ratings for each asset using all ratings present in the dataset. For those rating algorithms that contain a random rater selection component, *Random* and ϵ-first, we run each rating experiment 10 times using a different random selection of raters and report the average performance.

3. Results and discussion

We performed the evaluation for varying values of N. We found that the MAB approaches perform better in all cases than random rater selection. To illustrate this, the results using seven raters ($N = 7$) are presented in Fig. 1 which shows the overall performance as each asset is rated by the selected raters.

We also found that, in general, IEThresh fails to reach the performance of the other MAB techniques and always uses more raters, approximately ten at each step for MovieLens 10M and 13 for Jester as compared with the seven used for the other techniques shown in Fig. 1. The performance of the Overall-Best approach over the MAB approaches however, shows that there is still some room for improvement. It is also worth noting that for the task in hand the performance of the relatively simple MAB technique, the ϵ-first strategy, is higher than that of the much more sophisticated KL-UCB algorithm.

Overall, our results present strong evidence for the suitability of MAB approaches to the task of dynamic estimation of rater reliabilities and suggest that additional research in this direction is worthwhile.

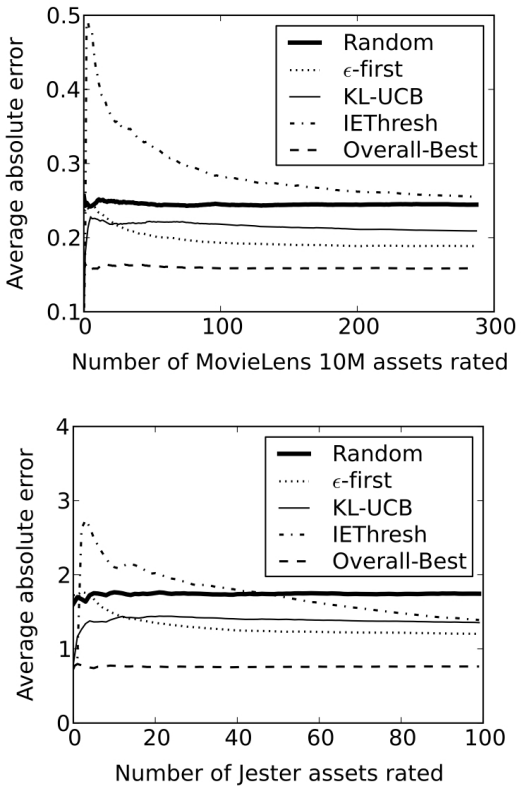

Figure 1. Using the top seven raters on each dataset

Acknowledgements

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

References

Ambati, V., Vogel, S., and Carbonell, J. Active Learning and Crowd-sourcing for Machine Translation. In *Procs of LREC'10*, pp. 2169–2174, 2010.

Brew, A., Greene, D., and Cunningham, P. Is it Over Yet? Learning to Recognize Good News in Financial Media. Technical report, University College Dublin, 2010.

Caelen, O. and Bontempi, G. A Dynamic Programming Strategy to Balance Exploration and Exploitation in the Bandit Problem. *Annals of Mathematics and Artificial Intelligence*, 60(1-2):3–24, 2010.

Donmez, P., Carbonell, J.G., and Schneider, J. Efficiently Learning the Accuracy of Labeling Sources for Selective Sampling. In *Procs of SIGKDD*, pp. 259–268, 2009.

Garivier, A. and Cappé, O. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In *Procs of COLT*, pp. 1–14, 2011.

Maron, O. and Moore, A.W. Hoeffding Races: Accelerating Model Selection Search for Classification and Function Approximation. In *Procs of NIPS*, pp. 59–66, 1994.

Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., and Moy, L. Learning From Crowds. *JMLR*, 11:1297–1322, 2010.

Sheng, V.S. and Provost, F. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers Categories and Subject Descriptors. In *Procs of KDD*, pp. 614–622, 2008.

Vermorel, J. and Mohri, Mehryar. Multi-Armed Bandit Algorithms and Empirical Evaluation. In *Procs of ECML*, pp. 437–448, 2005.

Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. *Advances in Neural Information Processing Systems*, 22(1):2035–2043, 2009.