International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IJIEMR Transactions, online available on 6th Mar 2018. Link :http://www.ijiemr.org/downloads.php?vol=Volume-08&issue=ISSUE-03

Title: **CROWD SOURCING CLASSIFICATION BY USING DLTA FRAMEWORK**

Volume 08, Issue 03, Pages: 18–22.

Paper Authors

MR.K.SRINIVASA RAO, I.MALLIKARJUNA REDDY

Vignan's Lara Institute of Technology & Science

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

CROWD SOURCING CLASSIFICATION BY USING DLTA FRAMEWORK

MR.K.SRINIVASA RAO[1], I.MALLIKARJUNA REDDY[2]
Assistant Professor[1], Department of M.C.A, Vignan's Lara Institute of Technology & Science
M.C.A Student[2] Department of M.C.A ‚Vignan's Lara Institute of Technology & Science

Abstract:

The increasing popularity of crowd sourcing markets enables the application of crowd sourcing classification tasks. How to conduct quality control in such an application to achieve accurate classification results from noisy workers is an important and challenging task, and has drawn broad research interests. However, most existing works do not exploit the label acquisition phase, which results in their disability of making a proper budget allocation. Moreover, some works impractically make the assumption of managing workers, which is not supported by common crowd sourcing platforms such as AMT or Crowd Flower. To overcome these drawbacks, in this paper, we devise a DLTA (Dynamic Label Acquisition and Answer Aggregation) framework for crowd sourcing classification tasks. The framework proceeds in a sequence of rounds, adaptively conducting label inference and label acquisition. In each round, it analyzes the collected answers of previous rounds to perform proper budget allocation, and then issues the resultant query to the crowd. To support DLTA, we propose a generative model for the collection of labels, and correspondingly strategies for label inference and budget allocation. Experimental results show that compared with existing methods, DLTA obtains competitive accuracy in the binary case. Besides, its extended version, which plugs in the state-of-the-art inference technique, achieves the highest accuracy.

Introduction

The rise of crowdsourcing markets such as AMT (Amazon Mechanical Turk) [1] enables the utilization of crowd powers to solve human intelligence tasks. Many classification tasks (e.g., image annotation, sentiment analysis, and website classification) are posted on such platforms to attain labels. Given a set of items and their potential categories (labels), a classification task aims to recognize the true label for each item. For example, given a set of facial images, one may request the crowd to classify them into 'smile', 'crying' and 'angry' according to the facial expressions. Apparently utilizing the crowd to perform such tasks is a wise choice, as personal inspection can be quite upset and burdensome when the number of items is large. However, due to the various expertise and reliability of workers, the labels provided by workers may have errors [30]. Acquiring multiple answers for each item and performing aggregation is a common practice to deal with such errors. According to the experiments on real crowdsourcing data [26], the simple majority voting

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

decision method has its limits on the accuracy. Therefore, how to obtain high-quality classification results through crowdsourcing is still under a great concern. To achieve high-accuracy classification results from crowd, system-oriented approaches have been widely studied. System works [10], [3], [36] set up various system mechanisms to perform effective quality control. The problem lies in their setting that the system can determine which tasks are assigned to which workers. This is not possible on AMT and CrowdFlower where tasks are self-selected by the workers. The application scenario concerned in this paper is to help task requesters of common crowdsourcing markets (AMT and Crowd-Flower) to improve the classification accuracy. Therefore, instead of following these system works, in this paper, we focus on solutions that can be adopted by users of common platforms like AMT and CrowdFlower. There are basically two types of such solutions proposed so far. The first type of solutions is the pure inference methods [32], [7], [24], [33]. The input and output of these pure inference methods are the crowdsourced labels and the inferred labels respectively. They have not exploited the budget allocation issue in the label acquisition phase, and just conduct uniform budget distribution. The second kind of solutions [4], [2] considers both label inference and label acquisition. With parameters estimated in the inference phase, they allocate budget to tough items in the label acquisition phase. Benefited from a more effective usage of budget, these adaptive methods generally achieve a higher accuracy than pure inference methods. However, their drawback lies in the assumption of managing workers. In this assumption, the requesters can target a specific worker to answer questions. In other words, the requesters are able to assign tasks to specific workers. This is not possible on typical crowdsourcing markets like AMT or CrowdFlower, as stated in [18]. On common crowdsourcing platforms AMT and Crowd-Flower, requesters decide which tasks to ask, and the posted tasks are self-selected by the workers. In pursuit of a higher classification accuracy using common platforms, we build a classification crowd sourcing framework in this paper, called Dynamic Label Acquisition and Answer Aggregation (DLTA, T means triple) framework. As shown in Figure 1, the DLTA framework follows an adaptive design for humanmachine- crowd interactions. In this framework, the machine serves as the medium between the user and the crowd. It proceeds in a sequence of rounds, and each round consists of two steps, the **label inference** (answer aggregation) step and the **label acquisition** step. In the label inference step, after the previous issued query is completed by the crowd, it conducts inference with new answers and then outputs its estimation to help the user decide the amount of budget for next round. The budget is given dynamically in a round manner in real-time by the user. In the label acquisition step, given the budget for the next round by the requester, the machine allocates the budget among items, and then issues the corresponding query to the crowd through platforms like AMT. Note that in the label acquisition step, the requester can alternatively choose to stop whenhe/she is satisfied with the present output.

Existing system:

The increasing popularity of crowd sourcing markets enables the application of crowd sourcing classification tasks. How to conduct quality control in such an application to achieve accurate classification results from noisy workers is an important and challenging task, and has drawn broad research interests. However, most existing works do not exploit the label acquisition phase, which results in their disability of making a proper budget allocation.

Proposed system

we devise a DLTA (Dynamic Label Acquisition and Answer Aggregation) framework for crowd sourcing classification tasks. The framework proceeds in a sequence of rounds, adaptively conducting label inference and label acquisition. In each round, it analyzes the collected answers of previous rounds to perform proper budget allocation, and then issues the resultant query to the crowd. To support DLTA, we propose a generative model for the collection of labels, and correspondingly strategies for label inference and budget allocation. Experimental results show that compared with existing methods, DLTA obtains competitive accuracy in the binary case. Besides, its extended version, which plugs in the state-of-the-art inference technique, achieves the highest accuracy.

Modules:

Password Policy:-

Our system imposes the following less restricted and practical to implement conditions on password selection.

1) Username or its sub-string should not appear in the password.

2) The password should contain at least 8 characters including alphabets, special symbols and digits.

Typo-Safety:

Very high probability by maintaining a minimum distance between the password and each generated honey word. We suggest using 'Levenshtein distance' to compute the distance between password and honey words. 'Lavenshtein distance' is calculated by counting the number of deletions, insertions, or substitutions required to transform one string into another. It can be used to calculate distances between variable length strings. In this way, all types of human typing errors can be taken into account. Legacy-UI password changes:

Evolving password model:-

Define the key terms for the better understanding of the scheme. These terms are defined with respect to the available disclosed password databases.

Token: We consider token as a sequence of characters that can be treated as a single logical entity. In our context, for a given password, tokens are the alphabet-strings (A), digit-strings (D) or special-character-strings(S).

Pattern: The different combinations of tokens form patterns for a password, e.g., ADS1, AS2D, S1AS1D etc. Note: To create honey words indistinguishable from user

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

password we do not preserve length of alphabets and digits, however we preserve the length of special-characters. Therefore the length of the special-character is mentioned as a subscript of S in the representation of pattern.

Frequency: It is the number of occurrences of the tokens or the password pattern in the available password database.

User-profile model: It generates honey words by combining some details from the user profile and checks the threshold of minimum distance with the password.

i) Create separate list of tokens named, token digits, token alphabet, token special Char from the information provided in user profile.

ii) To create k honey words, take k different combinations of elements from each token lists, satisfying the password policy of the service.

iii) Compare the tokens of the password with the tokens of the honey word. Reject the honey word if more than one token matches with password.

Conclusion:

In this paper, we propose DLTA, a framework for dynamic crowd sourcing classification tasks. It overcomes the drawbacks of existing works and is general and flexible. To implement the framework, we devise a joint probabilistic model for the generation of crowd sourced labels. Subsequently, techniques for label inference and budget allocation are deducted and presented. More specifically, we adopt the EM approach to estimate the model's parameters and the posterior probabilities of items' ground truth. Then, we formulate the budget allocation problem in the label acquisition phase, and propose the round-robin based Greedy algorithm as a solution. Finally, by comparison with the state-of-the-art methods, experimental results on both real and synthetic data demonstrate the significance of our method.

References

1.Amazon mechanical turk. https://www.mturk.com/mturk/ welcome.

2.Y. Bachrach, T. Graepel, T. Minka, and J. Guiver. How to grade a test without knowing the answers—a bayesian graphical modelfor adaptive crowdsourcing and aptitude testing. arXiv:1206.6386, 2012.

3.R. Boim, O. Greenshpan, T. Milo, S. Novgorodov, N. Polyzotis, and W.-C. Tan. Asking the right questions in crowd data sourcing. In Proc. of the ICDE, pages 1261–1264. IEEE, 2012.

4.X. Chen, Q. Lin, and D. Zhou. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In Proc.of the ICML, pages 64–72, 2013.

5. N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi. Aggregating crowdsourced binary ratings. In Proceedings of the 22nd international conference on World Wide Web, pages 285–294. ACM, 2013.

6.A. Das Sarma, A. Parameswaran, and J. Widom. Towards globally optimal crowdsourcing quality management: The uniform worker setting. In Proc. of the SIGMOD, pages 47–62. ACM, 2016.

7.A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. Applied statistics, pages 20–28, 1979.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

8.A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological), pages 1–38, 1977.
9. P. Donmez, J. G. Carbonell, and J. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In Proc. of the SIGKDD, pages 259–268, 2009.
10. J. Fan, G. Li, B. C. Ooi, K.-l. Tan, and J. Feng. icrowd: An adaptive crowdsourcing framework. In Proc. of the SIGMOD, pages 1015– 1030. ACM, 2015.
11. S. Faradani, B. Hartmann, and P. G. Ipeirotis. What's the right price? pricing tasks for finishing on time. Human computation,11:11, 2011.
12. M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In Proc. of the SIGMOD, pages 61–72. ACM, 2011.
13. S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence, (6):721–741, 1984.
14. D. Haas, J. Wang, E. Wu, and M. J. Franklin. Clamshell: Speeding up crowds for low-latency data labeling. Proc. of the VLDB,9(4):372–383, 2015.
15. C.-J. Ho, S. Jabbari, and J. W. Vaughan. Adaptive task assignment for crowdsourced classification. In Proc. of the ICML, pages 534–542, 2013.
16. N. Q. V. Hung, N. T. Tam, L. N. Tran, and K. Aberer. An evaluation of aggregation techniques in crowdsourcing. In Proc.of the WISE, pages 1–15, 2013.
17 P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In Proc. of the SIGKDD workshop on human computation, pages 64–67, 2010.
18. D. R. Karger, S. Oh, and D. Shah. Budget-optimal task allocation for reliable crowdsourcing systems. Operations Research, 62(1):1– 24, 2014.
19. F. Laws, C. Scheible, and H. Sch ¨ utze. Active learning with amazon mechanical turk. In Proc. of the EMNLP, pages 1546–1556, 2011.
20. D. D. Lewis. Evaluating and optimizing autonomous text classification systems. In Proc. of the SIGIR, pages 246–254. ACM,1995.

15. Bibliography

(1) Java Complete Reference by Herbert Shield
(2) Database Programming with JDBC and Java by George Reese
(3) Java and XML By Brett McLaughlin
(4) Wikipedia, URL: http://www.wikipedia.org.
(5) Answers.com, Online Dictionary, Encyclopedia and much more, URL: http://www.answers.com
(6) Google, URL: http://www.google.co.in
(7)Project Management URL: http://www.startwright.com/project.html