Journal of Cyber Security
DOI:10.32604/jcs.2020.07229
Article

Preserving the Efficiency and Quality of Contributed Data in MCS via User and Task Profiling

Dingwen Wang and Ming Zhao*

School of Computer Science and Engineering, Central South University, Changsha, 410082, China
*Corresponding Author: Ming Zhao. Email: meanzhao@csu.edu.cn
Received: 09 May 2019; Accepted: 09 June 2020

Abstract: Mobile crowdsensing is a new paradigm with powerful performance for data collection through a large number of smart devices. It is essential to obtain high quality data in crowdsensing campaign. Most of the existing specs ignore users' diversity, focus on solving complicated optimization problem, and consider devices as instances of intelligent software agents which can make reasonable choices on behalf of users. Thus, the efficiency and quality of contributed data cannot be preserved simultaneously. In this paper, we propose a new scheme for improving the quality of contributed data, which recommends tasks to users based on calculated score that jointly take the matching degree and task's rationality into account. We design QIM as Quality Investigation Mechanism for profiling tasks' rationality and matching degree, which draw on support vector machine (SVM) to learn it from historical data. Our mechanism is validated against the examination in experiment, and the evaluation demonstrates that the QIM mechanism achieves a better performance while improving efficiency E and quality Q at the same time compared with benchmarks.

Keywords: Crowdsensing; matching degree; support vector machine

1 Introduction

The core of the crowdsensing campaign [1] is obtaining data, which combines the idea of crowdsourcing and initiative of the group to achieving data. Different from the traditional labor market which assigns job to designated workers, sensing tasks are displayed on the software platform and the contributor could be recruited online. A typical crowdsensing campaign usually consists of three components: the MCS server, the tasks and participants. The MCS server is responsible for all costs in campaign process, and the participants contribute data by completing tasks. For contributors, in the light of that there is a price to carry out tasks, such as spending energy and submitting privacy, incentives are necessary for providing certain compensation [2] to maintain the enthusiasm of contributor. Driven by the promising reward, the power of public grants crowdsensing leads a huge advantage compared with other paradigms, it could apply in various fields to provide services, such as traffic prediction [3], environmental information collection [4], spectrum analysis [5], and so on.

With regard to the efficiency and quality of collected data, the main mind is to design recommendation model. There exist two classical methods. The first one is user-centric model, in which the task is first filtered then displayed according to user's attributes [6] Essentially, this model is still based on the user's own decision, despite the task is preprocessed. Another method is platform-centric model. The recommended task is determined by historical record and the expectation that submitted by the user [7]. But this assumption ignores the complexity of user behaviors and could be oversimplification of property. On this account, we use the efficiency E and quality Q for measuring the whole campaign. We define that E is the efficiency about data collection, which is measured by the acceptance ratio. And Q is the quality of collected data which is measured by the sum of entire difference between the contributed data and the ground truth.

For the aim of improving efficiency and quality of contributed data, we prepare to expand a new scheme build on platform-centric model, which focuses on tasks' rationality and facilitates the matching degree of the users with appropriate tasks in the meantime. Considering the limitations of existing model, we jointly consider the profiling of user and attributes of the task in MCS, and propose a precisely recommended framework. Concretely, different from the traditional task's features that only have two attributes (the distance and payment), we collect multiple attributes and take the time into a significant position. As for the issue of matching degree, the machine learning [8] is a good solution, in that the trained simulator could model users' thought, approximately [9].

There exist an abundance of research works on data [10] collection for mobile crowdsensing [11]. And the provision of incentives has been looked upon in a longer time. In [12], for service quality of data collection, a new system model is proposed based on reverse auction framework, which formulate the problem as the social optimization user selection (SOUS) problem. Considering the different cases in reality, the authors designed two incentive mechanisms, MST and MMT, to select users, which is optimal algorithm based on dynamic programming in single time window and optimal solution based on greedy approach in multiple time window. In [13], the authors design RIT as a robust incentive tree mechanism, which combines the advantages of auction with incentive tree to motivate users for participation and solicitation. In [14], for attracting more user participation, the authors designed incentive mechanisms and considered two system models: the crowdsource-centric model and the user-centric model. Although these works have designed efficient incentives for data collection from different aspects, none of them take user's quality and matching degree into account.

Apart from incentive mechanism [15–17], the problem of user's reliability and preference [18] has been widely studied to handle the issue of data collection too. In [19], the authors propose a novel crowd-based credibility improving scheme (CCIS) to enhance the credibility of the data, through the lightweight fixed-width clustering algorithm, and accompany the reputation knowledge, the scheme could identify the erroneous data and ensure a certain quality of information. Besides, without the authenticity judged by reputation knowledge [20], the authors combined the data clustering and logical reasoning to create a new scheme named CLOR to reach the effect of distinguishing error information. Whereas they considered the quality of contributed data, which are based on user-centric model, and incapable of recommending personalized tasks to users.

2 Model and Problem Statement

It is difficult to strike a balance between efficiency E and quality Q during the process of sensing data collection, we tend that E is measured by the ratio that the number of all completed tasks divided by the total recommended tasks to users and Q is measured by the result that the sum of the entire difference between the data that contributed by users and the ground truth. We design QIM as a Quality and Matching Investigation Mechanism for mobile crowdsensing. Before the task t is offered to user u, we will optimize them through the QIM, the specific steps as followed.

$$S_{\mathrm{ij}}=\alpha \cdot S_{\mathrm{ui}}+\beta \cdot S_{\mathrm{ui\text{-}tj}}+\gamma \cdot S_{\mathrm{tj}} \tag{1}$$

α and β are weight parameters, which emphasize the proportion of corresponding part. We can see that the key is to estimate the matching degree and task's rationality. In this paper, the core idea of obtaining is to calculate the ratio between hindrance and driven factors. Therefore, with the as small as possible, the task that submitted by is more credible, and the final efficiency E and quality Q is better. In Next, we will present the specific forms of them.

2.1 Task's Rationality

In our campaign, we set that the task should be performed immediately if they were accepted. For the rationality of tasks, we follow the rule that whether the task is completed or not which is influenced by the reasonable of itself. In our image, the task feature vector is consisted of multiple attributes, which has same dimensions and different specific feature values.

From the perspective of convenience and integrity, we decide that the task feature vector is comprised of 5 attributes, F=($\mathrm{F_t}$, $\mathrm{F_l}$, $\mathrm{F_p}$, $\mathrm{F_T}$, $\mathrm{F_m}$). Specifically, $\mathrm{F_t}$ is symbol of tasks' type, $\mathrm{F_l}$ is physical location of the task, $\mathrm{F_p}$ is payment for task, $\mathrm{F_T}$ is the time cost and $\mathrm{F_m}$ is the moment at this time. Then, we formalize the task's rationality as follows.

$$S_{\mathrm{tj}} = \frac{F_T}{F_p} e^{F_m} \tag{2}$$

The main thought of the above formula is comparing drivers with blocks for obtaining thetask's rationality score. In Eq. (2), Where the mainly control the congestion parameter, in that the user is not free in every moment of a day and the traffic congestion are also variety. Due to the range of feature vector value is varies, we preprocessing the data, scale it into the same range and then to calculate the target score by Eq. (2).

2.2 Matching Degree

For the matching degree between user and task, we treat as a classification problem, extract attributes and draw on machine learning algorithm to train it. The modified learning optimization for the S_{ui-tj} is given by

$$S_{ui-tj} = \exp(-(Class)) \tag{3}$$

In Eq. (3) where the $Class$ is the label of classification, users could take it or leave it when a task is recommended by platform. In here, we set that the identifier -1(C_{-1}) represents refusing action and the identifier +1(C_{+1}) represents accepting action. The process about decision making is modeled by support vector machine. And the data set that trained for it is composed of real data and synthetic data.

we defined that the features of dataset U-T are constructed by user's and task's features. The task's features as showing in above, and the user's features include the physical location where the user right now, the expectation of reward, the interest about what they prefer to carry out. We formalized that $\mathrm{U}-\mathrm{T}=\{(x_{1n},y_1),(x_{2n},y_2),\ldots,(x_{nn},y_n)\}$, in (x_{1n},y_1), where x_{1n} represents n attributes and y_1 means the class value equals to -1 or +1, which depend on the decision made by user. For the purpose of figuring out how to classify the class value, we need to learn the linear equations given as follows of the separated hyperplane through data set U-T.

$$w^T * x + b = 0 \tag{4}$$

Due to the set U-T is multidimensional and manifest itself has nonlinear separable characteristic, the kernel function is absolute. We use the Radial Basis Function (RBF) to process the data and the decision function given as

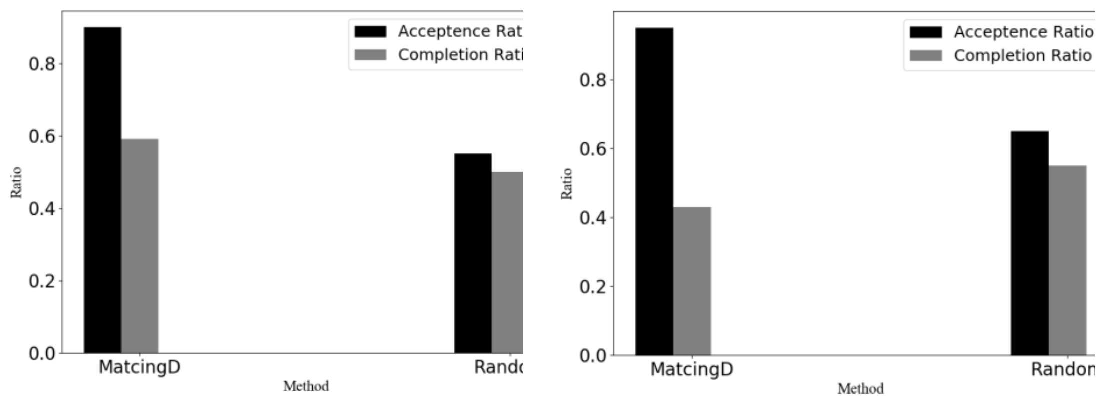
$$f(x) = sign(\sum_{i=1}^{N} a_i^* y_i \exp(-\frac{\left\| x_i - x_j \right\|^2}{2\sigma^2}) + b) \tag{5}$$

We omit the explanation of letter, in that the Eqs. (4) and (5) is classical formulation for SVM, and commonly be cited in this industry. In processing of learning decision boundary, we should regard the regularization penalty λ and parameter gamma as very significant for preventing overfitting and over-confident predictions. Therefore, we achieve the λ and gamma by cross-validation, and determined the final model M_{svm} based on accuracy of test set.

3 Evaluation

We examine the above model with samples to confirm the effect of our proposed method. In experiment, we recruit twelve users and set 100 tasks in 5 categories in advance, collect the ground truth (which is unauthorized) for users by ourselves to make sure the credit. All tasks are showed in mobile terminal, users can browse and choose tasks depend on themselves. We compare the operation records with ground truth to conduct conclusion.

We evaluate our proposed method by efficiency E and quality Q In experiment. More specifically, which means that we adopt the task acceptance rate (TAR) and data accuracy rate (DAR) to lend concreteness to the E and Q. TAR is calculated by the number of completed tasks divided by total recommended tasks, and DAR is the sum of distance between submitted data and ground truth. DAR should be treated separately due to the existence of two data types–continuous data and numerical data. For continuous data, we use Root Mean Square Error (RMSE) to quantify the DAR, the mathematical expression of RMSE is defined as $\sqrt{\sum_{j\in T}(d_{ij} - d_j^g)^2/N_t}$. And for categorical data, we use Error Rate to measure the DAR, mathematically, the ER is defined as $\sum_{j\in T} L(d_{ij}, d_j^g)/N_t$.



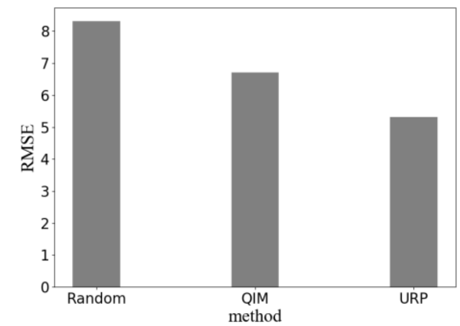
(a) (b)

Figure 1: Comparison of different matching strategies

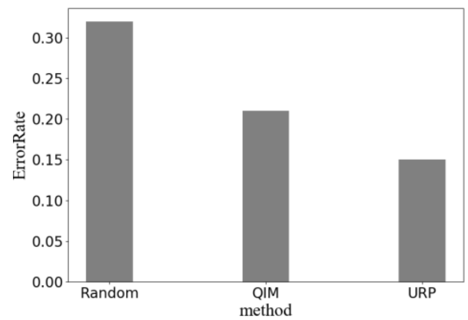
The comparative performance of two methods is presented in Fig. 1. for acceptance ratio, we can see that the matching degree measured method is more outstanding than random method. However, in contrast to acceptance ratio, the difference of completion ratio between two methods could be negligible. The main reason is that the matchingtasks without constraints of reliability is not guaranteed.

Now, to evaluate the performance of our proposed paradigm, we recommend 20 tasks to each user and adopt two benchmarks, the random recommendation and user reliability- preference recommendation (URP) [21]. The random recommendation provides users 20 tasks randomly, the URP paradigm provides 20 tasks with the highest probability or score to each user. For our mechanism, the parameters and in Eq. (1) is set to 0.4 and 0.6, respectively, which provide each user 20 recommended tasks with the highest score that calculated by QIM mechanism. We demand users to choose tasks, estimate TAR and DAR that computed by RMSE and Error Rate and then present as graphics.

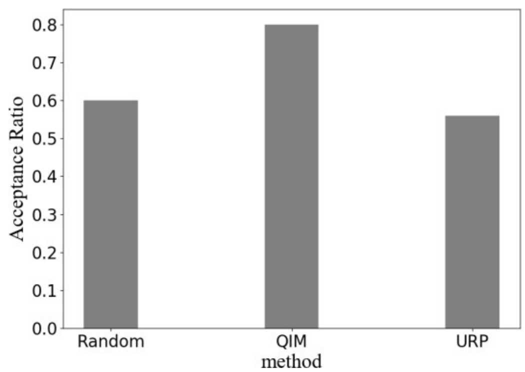
In Fig. 2(a), we can find that the PP paradigm has the highest acceptance ratio, which is preference-origin method and it fit with the analysis we conclude as shown in above. While as for the accuracy rate presentation, the URP strategy acquires the best results as showing in Figs. 2(b) and 2(c), in that the URP method is mainly concentrated on reliability of users. Although our paradigm is not the best compared with other in a single field, which could maximize the efficiency E (TAR) and quality Q (DAR) at the same time.



(a)



(b)



(c)

Figure 2: Comparison of acceptance and accuracy with different strategies

4 Conclusion

In this paper, we have studied recommendation model in mobile crowdsensing, concentrate on personalized task and contributed data quality with time attribute. We design QIM as Quality Investigation Mechanism that can recommend tasks to users based on quality and matching degree. The quality algorithm originates from task's rationality and matching degree, we focus on task itself for task's rationality and by drawing on support vector machine to address matching degree. Finally, with the help of volunteers, the evaluation has been conducted to evaluate our framework with other benchmarks. And the performance confirms that the QIM can improve efficiency E and quality Q at the same time.

Funding Statement: This work was supported in part by the National Science Foundation of China under Grant 61572526.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. K. Ganti, F. Ye and H Lei, "Mobile crowdsensing: current state and future challenges," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 32–39, 2011.

[2] C. M. Chou, K. C. Lan and C. F Yang, "Using virtual credits to provide incentives for vehicle communication," in *12th International Conference on ITS Telecommunications*, Taipei, pp. 579–583, 2012.

[3] X. Wang, J. Zhang, X. Tian, X. Gan, Y. Guan *et al.*, "Crowdsensing-based consensus incident report for road traffic acquisition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 8, pp. 2536–2547, 2018.

[4] X. Kang, L. Liu and H. Ma, "Data correlation based crowdsensing enhancement for environment monitoring,"

in *IEEE International Conference on Communications, Kuala Lumpur*, pp. 1–6, 2016.

[5] I. F. Akyildiz, B. F. Lo and R. Balakrishnan, "Cooperative spectrum sensing in cognitive radio networks," A survey. *Physical Communication*, vol. 4, no. 1, pp. 40–62, 2011.

[6] S. Bradai, S. Khemakhem and M. Jmaiel, "Re-OPSEC, Real time opportunistic scheduler framework for energy aware mobile crowdsensing," in *24th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, Split, pp. 1–5, 2016.

[7] K. Han, E. A. Graham, D. Vassallo and D. Estrin, "Enhancing motivation in a mobile participatory sensing project through gaming," in *IEEE Third International Conference on Privacy, Security, Risk and Trust and IEEE Third International Conference on Social Computing*, Boston, MA, pp. 1443−1448, 2011.

[8] L. Xiao, Y. D. Li, G. A. Han and H. Dai, "A secure mobile crowdsensing game with deep reinforcement learning," *IEEE Transactions on Information Forensics & Security*, vol. 13, no. 1, pp. 35-47, 2017.

[9] J. X. Zhang, N. Xie, X. J. Zhang, K. Yue, W. D. Li *et al.*, "Machine learning based resource allocation of cloud computing in auction," *Computers, Materials & Continua*, vol. 56, no. 1, pp. 123−135, 2018.

[10] J. R. Cheng, R. M. Xu, X. Y. Tang, V. S. Sheng and C. T Cai, "An abnormal network flow feature sequence prediction approach for DDoS attacks detection in big data environment," *Computers, Materials & Continua*, vol. 55, no. 1, pp. 95−119, 2018.

[11] Y. Yang, Y. W. Chen, Y. L. Chen and W. Bi, "A novel universal steganalysis algorithm based on the IQM and the SRM," *Computers, Materials & Continua*, vol. 56, no. 2, pp. 261–272, 2018.

[12] J. Xu, J. Xiang and D. Yang, "Incentive mechanisms for time window dependent tasks in mobile crowdsensing," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6353−6364, 2015.

[13] X. Zhang, G. Xue, R. Yu, D. Yang and J. Tang, "Robust incentive tree design for mobile crowdsensing," in *IEEE 37th International Conference on Distributed Computing Systems*, Atlanta, GA, pp. 458−468, 2017.

[14] D. J. Yang, G. L. Xue, X. Fang and J. Tang, "Incentive mechanisms for crowdsensing: crowdsourcing with smartphones," *IEEE/ACM Transactions on Networking*, vol. 24, no. 3, pp. 1732−1744, 2016.

[15] B. Faltings, J. J. Li and Jurca, R. "Incentive mechanisms for community sensing," *IEEE Transactions on Computers*, vol. 63, no. 1, pp. 115−128, 2014.

[16] H. Jin, L. Su, D. Chen, K. Nahrstedt and J. Xu, "Quality of information aware incentive mechanisms for mobile crowd sensing systems," *MobiHoc*, 2015.

[17] Y. Ueyama, M. Tamai, Y. Arakawa and K. Yasumoto, "Gamification-based in-centive mechanism for participatory sensing," in *IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, Budapest, pp. 98−103, 2014.

[18] M. Karaliopoulos, I. Koutsopoulos and M. Titsias, "First learn then earn: optimizing mobile crowdsensing campaigns through data-driven user profiling," *MobiHoc*, 2016.

[19] T. Q. Zhou, Z. P. Cai, M. Xu and Y. Y. Chen, "Leveraging crowd to improve data credibility for mobile crowdsensing," *Computers & Communication, IEEE*, 2016.

[20] G. Bigwood and T. Henderson, "IRONMAN: using social networks to add incentives and reputation to opportunistic networks," in *IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, Boston, MA, pp. 65−72, 2011.

[21] S. Yang, K. Han, Z. Zheng, S. Tang and F. Wu, "Towards personalized task matching in mobile crowdsensing via fine-grained user profiling," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, Honolulu, HI, pp. 2411−2419, 2018.