A Posted Pricing Mechanism Based on Random Forests in Crowdsourcing Market

Lifei Hao[1,2], Bing Jia[1,2(✉)], and Chuxuan Zhang[1,2]

[1] College of Computer Science, Inner Mongolia University, Hohhot 010021, China
jiabing@imu.edu.cn

[2] Inner Mongolia A.R. Key Laboratory of Wireless Networking and Mobile Computing, Hohhot 010021, China

Abstract. With the rapid development of the Internet, the combination of outsourcing and Internet has produced an overturning mode for labor cooperation – crowdsourcing. Crowdsourcing outsource the work that used to be done by internal staffs of a company or organization to non-specific people in a free and voluntary way, which concentrates the wisdom of public to solve difficult problems, greatly optimizes the rational allocation of human resources and thus improves the social productivity. In the environment of crowdsourcing market, how to set an "appropriate" price to recruit workers to complete a given task at a reasonable quality and cost is a key problem which restricts the development of it. Therefore, this paper proposes a posted pricing method based on the Random Forests (RF) algorithm in crowdsourcing market. The proposed mechanism is described theoretically and the actual crowdsourcing date is acquired from Taskcn by python spider firstly. Then, based on these empirical data, serval typical machine learning methods have been compared, which proves that RF is a very suitable method for posted pricing in crowdsourcing market. Finally, extensive experiments have been conducted and analysed for optimizing the parameters in RF and a set of parameters suitable for posted pricing in crowdsourcing is given to construct the corresponding RF model.

Keywords: Crowdsourcing · Pricing mechanism · Random Forests · Machine learning · Web spider

1 Introduction

In recent years, as the rapid development of Internet technology and combing with outsourcing, a innovative mode of labor cooperation – crowdsourcing [7] has emerged, which means that the work previously performed by employees in a company or institution can be outsourced to non-specific mass groups in a free and voluntary way through Internet technology. In practice, whether in domestic or abroad, there are many crowdsourcing platforms, thus the crowdsourcing

J. Li et al. (Eds.): SPNCE 2019, LNICST 284, pp. 627–636, 2019.
https://doi.org/10.1007/978-3-030-21373-2_52

market formed. Currently, crowdsourcing has been applied on a large scale and there are still unlimited market opportunities to be explored by the further developing of it.

A typical crowdsourcing structure [13] universally consists of three subjects: the requester, the worker, and the crowdsourcing platform, as shown in Fig. 1. Requester is the initiator of a specific crowdsourcing activity, which mainly releases the task and provides rewards. Workers who complete tasks and receive rewards are the group composed of individuals or teams to solve problems. As an intermediary, the crowdsourcing platform hosts the rewards and connects requesters and workers. In the whole process of crowdsourcing, rewards is the most important incentive approach to enable workers completing tasks on time and with a high quality.

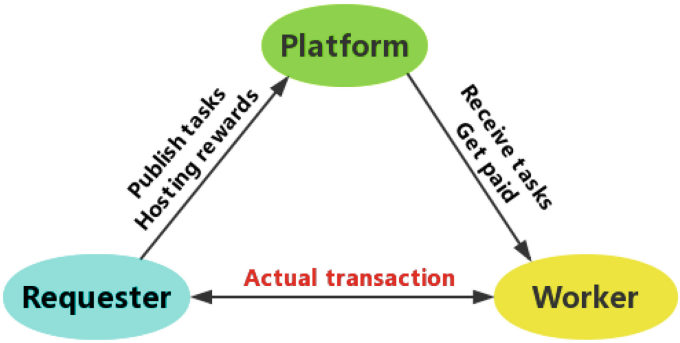


Fig. 1. The typical structure of crowdsourcing.

If the pricing for rewards is too low, the number of workers involved in the task will be insufficient, which will affect the completion of the task or make the quality lower. On the contrary, it will cause the waste of economic resources, increase the cost of employers and platforms, and thus reduce their benefits. Therefore, how to set an appropriate price to recruit groups to complete the task at a reasonable quality and cost, namely the crowdsourcing pricing problem, is the core issue restricting the development of crowdsourcing [4,5].

To solve the problem of posted pricing, [5] proposed a posted pricing mechanism based on quality-aware Bayesian model (assuming that the cost and completion quality of workers can be obtained from the probability distribution of historical data) in crowdsensing. [10] proposed a multi-parameter posted pricing mechanism based on collection-behavior by applying all-payment auction. [11] studied the using of autoregressive method with market sentiment indicators to predict American oil prices, and concluded that autoregressive method was not strong in predicting, so that the Machine Learning (ML) method should be considered to solve such problems. [8] converted the posted pricing problem into MAB (multi-armed bandit) problem to solve, and designed an optimal algorithm

to exploit the unique features of microtask crowdsourcing. This paper adopted the actual data of Amazon MTurk platform. Since a crowdsourcing platform will generate a large amount of historical data after running for a period of time, we can use ML to make empirical price prediction. Obviously, although each of the above methods proposed innovation to solve the posted pricing problem, ML method was not taken into account for such a typical price prediction problem (although mentioned in [11], not implemented). Therefore, this paper proposes a posted pricing mechanism based on the Random Forests (RF) [2] which belongs to an ensemble ML method in the crowdsourcing market.

The rest of paper is organized as follows. Section 2 discusses the theory of RF. Section 3 obtained the actual history transaction data in a crowdsourcing platform by using python spider technology. In Sect. 4, on the one hand, RF and other ML methods in price predicting are compared by means of data mining. On the other hand, RF is adjusted and optimized via experiments, and appropriate parameters are given. The final section summarizes the full paper and outlooks future work.

2 Random Forest

2.1 Basic Concepts

As a new rising, highly flexible ML algorithms, the Random Forests (RF) have a broad application prospects, e.g. it can be used to simulate marketing, analyze customer source, retention and loss, and also to predict the risk of disease as well as susceptibility. In recent years, the using of RF by participants in international and domestic competitions have been quite high.

RF's basic unit is Decision Tree (DT) and it is an algorithm that integrates multiple DTs by the idea of ensemble learning, a large branch of machine Learning. There are two key words in the name of RF. One is "random" and the other is "forests". "Forests" represents many DTs to be generated in the algorithm, while "random" represents random sampling and random selecting features. Intuitively, each DT is a regressor or classifier, so that n trees will have n regression or classification results for an input sample. The RF integrates all the regression results or voting results and calculates the mean value or specifies the category with the most votes as the final output.

For a DT in RF, in order to determine the order of feature selection, the concepts of Information (I), Entropy (E) and Information Gain (IG) need to be introduced. If there is a group of things waiting for classified that can be divided into a number of categories, the information of a certain class x_i can be defined by Formula (1). Where $p(x_i)$ is the probability when x_i occurs.

$$I(X = x_i) = -\log_2 p(x_i) \tag{1}$$

In information theory and probability theory, entropy is a measure of random variables' uncertainty. Combined with the information above, entropy is the

expected value of information, which can be denoted as follow.

$$E(X) = \sum_{i=1}^{n} p(x_i) I(x_i) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{2}$$

IG is an indicator used to select features in DT algorithm. The larger the IG is, the better the selectivity of this feature will be. In the probability theory, it is defined as the difference between the entropy of the set to be classified and the conditional entropy of a selected feature, as in the Formula (3).

$$IG(Y|X) = E(Y) - E(Y|X) \tag{3}$$

where

$$E(Y|X) = \sum_{x} p(x) E(Y|X = x) \tag{4}$$

2.2 The Algorithm

In the RF, the generation rules of each DT are as follows:

- **Step1** : For each DT, if the size of training set is n, n training samples are randomly and reversely (known as bootstrap sampling) extracted from the training set as the specific training set for this DT;
- **Step2** : If the dimensions of each sample are M, specify a constant $m < M$, and randomly select m features in M as subset. Each time the DT dividing, select the optimal one of the m features;
- **Step3** : Each DT grows to its maximum extent and there is no pruning process.

It can be seen that the training set of each DT is different and contains repeated training samples. Therefore, the classification effect (error rate) of RF is related to two factors: the greater the correlation between any two DTs in the forest, the greater the error rate; the stronger the classification ability of each DT in the forest, the lower the error rate of the whole forest. If the number of feature selection m is reduced, the relevance and classification ability of each DT will be reduced correspondingly and vice verse. So it is a key issue how to choose the optimal m. At the same time, increasing the number of DT in RF can improve the precision and regression results, but the algorithm efficiency will be reduced.

In addition, the Minimum Number of Sample Leaves (MNSL) of a DT, i.e. the minimum number of end nodes in a DT, will have an important impact on its prediction effect. This is because smaller MNSL will make it easier for the model to capture the noise in the data, which will result in the problem of over-fitting. In the same way, the maximum depth of DT can cause similar problems if not limited. Therefore, we will adjust and optimize these four parameters in the experiment (Sect. 4).

3 Data Acquisition

In this paper, the python spider [12] technology is used to acquire the historical transaction data from the crowdsourcing platform Taskcn [1], one of the earliest crowdsourcing platforms in China. The program flow of python spider used for this work is shown in Fig. 2. First, we have made a deep analysis to the source code of target site's webpage, identified URL format and the structure of each label. Secondly, the directory list of all completed tasks is traversed, and the URL to be visited is spliced by task ID. Then, regular expression and the third-party open source library Beautiful Soup [3] are used to parse the webpage and obtain the content of it. Finally, each transaction data acquired is processed into CSV format and stored on the hard disk. In addition, in the process of web page parsing, the spider also handle the issues such as messy code, task deleted, content confidential etc.

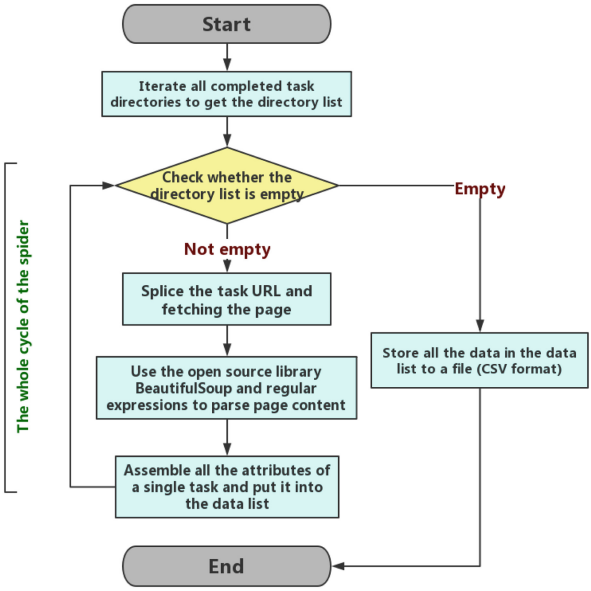


Fig. 2. The program flow chart of python spider.

Finally, after a period of crawling, 35,622 transaction data of all completed reward tasks on Taskcn from year 2006 to 2018 had been gotten. Each piece of data contains 18 attributes such as task ID, task category, requester's credit, task price etc. After data preprocessing such as data cleaning, useless attribute deleting and abnormal data eliminating, 34,805 pieces of valid data were finally obtained, and each had 11 available attributes.

4 Experiment

4.1 The Comparison of Machine Learning Methods

For the task data from the above section, each task contains multiple attributes (such as category, time limit, workload etc.). The ML methods can be used to predict its price by its attributes. Since the price is a continuous value, this is a typical regression problem, and we use the general process of data mining [9] to research.

The open source machine learning and data mining software Weka [6] had been used to compare the predicting results of some ML methods (e.g., M5 Model Tree (M5MT), Artificial Neural Network (ANN), k-Nearest Neighbour (KNN), Linear Regression (LR), etc.) that are suitable for dealing with regression problem. Mean Absolute Error (MAE) is used as the evaluation index, which defined as Formula (5). The verification method adopts k-fold cross validation and $k = 10$.

$$MAE = \frac{\sum_{i=1}^{n} AE_i}{n} \quad i \in \{1, 2, ..., n\} \tag{5}$$

where

$$AE_i = |actual_i - estimated_i| \quad i \in \{1, 2, ..., n\} \tag{6}$$

where n is the total number of data, $actual_i$ is the actual value of the i^{th} data in the test set, and $estimated_i$ is the estimated value.

As shown in Fig. 3, it can be seen that RF has almost the smallest MAE. In another words, compared with other ML methods, RF (and M5MT) has the best price prediction power with crowdsourcing data. Therefore, RF is a method that is very suitable for posted pricing in crowdsourcing market.

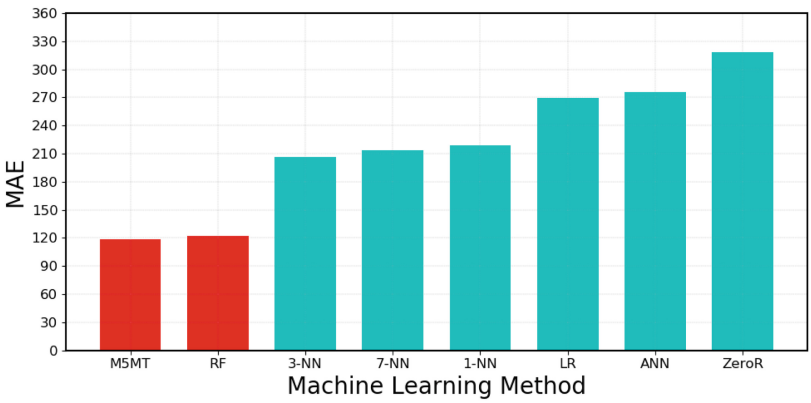


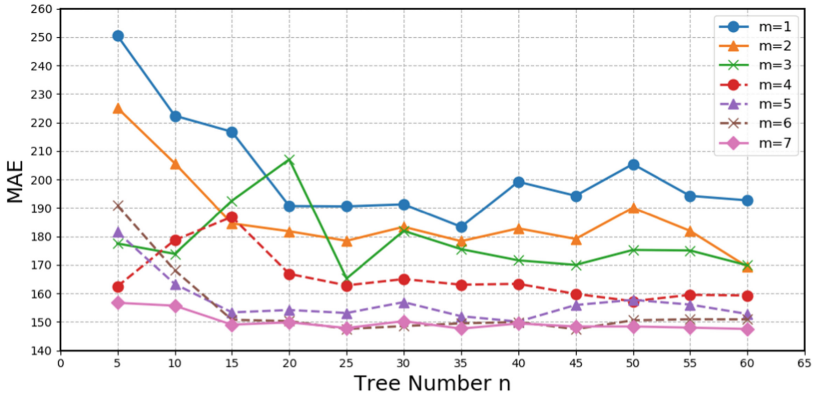
Fig. 3. The comparison between ML methods for price prediction in crowdsourcing.

4.2 Parameter Adjusting for Random Forests

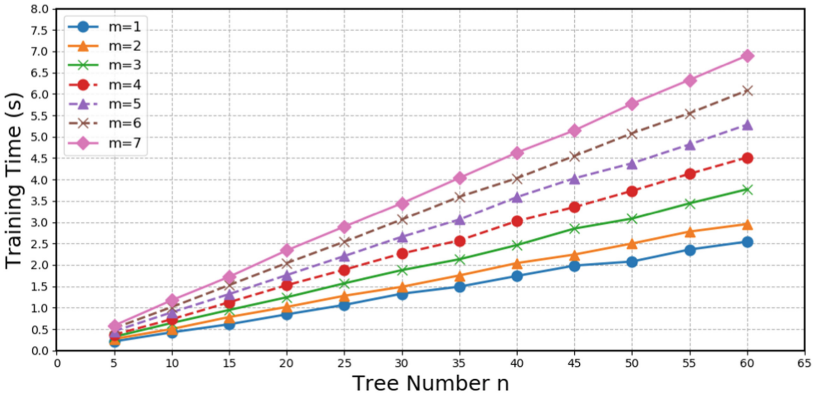
In order to study the influence of four parameters (tree number n, the number of feature selection m, the minimum number of sample leaves l, the maximum depth of each DT d) discussed in Sect. 2.2 on regression results, this section specially designed two experimental schemes, using MAE of k-fold cross validation and

Table 1. The parameter setting in two experiment schemes.

Parameters	Scheme-A	Scheme-B
n	$\{5, 10, 15, ..., 60\}$	40
m	$\{1, 2, ..., 7\}$	5
l	2	$\{10, 20, 30, 40\}$
d	∞	$\{3, 6, 9, ..., 30\}$



(a) MAE



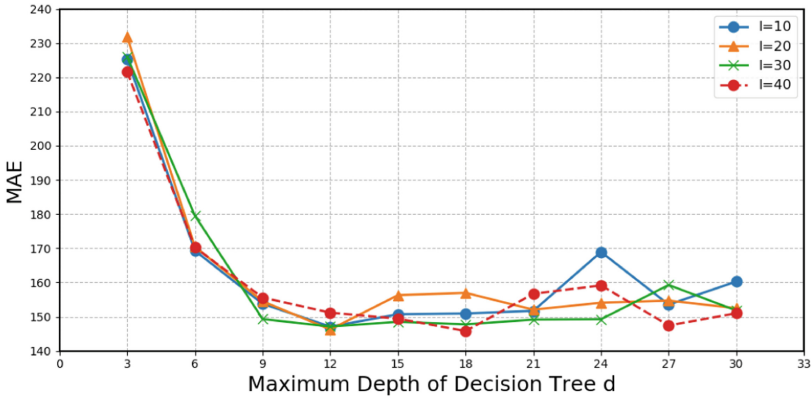
(b) Training Time

Fig. 4. The changing curve of MAE and training time as m and n increasing.

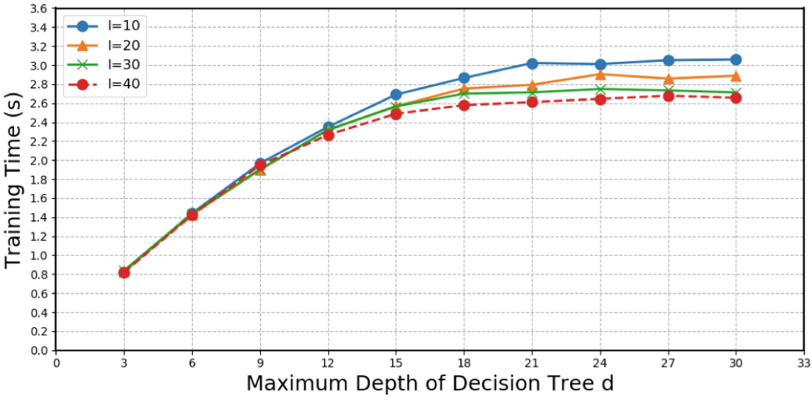
evaluating the training time of all data (34, 805). Both schemes are implemented by python on a PC with i5 CPU, 8 GB RAM and Win-7 (64 bit) operating system and designed via the control variable method, as shown in Table 1. In Scheme-A, $\{l, d\}$ are fixed values and $\{m, n\}$ changes, while in Scheme-B reversed.

When m increases gradually, MAE decreases (see Fig. 4(a)), while m increases to a certain value, MAE does not change significantly ($m = 5$, $m = 6$, $m = 7$ almost overlap). When m is large, the value of n has little influence on the prediction error. Both m and n have a linear effect on the training time (see Fig. 4(b)). Therefore, the prediction ability and training time of RF are contradictory corresponding to the setting of m and n. It should be tried to use smaller m and n when ensuring that MAE meets the requirements.

Then turn the lens to how l and d for a single DT affect the overall performance of RF. As can be seen from Fig. 5, when d is relatively small, the curve



(a) MAE



(b) Training Time

Fig. 5. The changing curve of MAE and training time as l and d increasing.

of MAE and training time almost overlapped, which is because l and m have a conflict. Due to the limiting for the maximum depth of DT, DT cannot be further divided, so only m works. When m becomes larger, l starts to act and thus the curve starts to differ. In Fig. 5(a), when m is about 10, MAE reaches the minimum. When $m < 10$, MAE curve is steep and training time curve is relative slow and MAE is not changing after $m > 10$. Besides, l has little effect on MAE and training time. In summary, $\{m, n, l, d\} = \{5, 40, 40, 10\}$ is a relative better setting of parameters in the case talked in this paper.

5 Conclusions

At present, the development speed of crowdsourcing is extremely rapid, from the traditional task crowdsourcing to the emerging mobile crowdsensing. However, rewards pricing as the most important incentive factor in crowdsourcing is the core issue that restricts its development. This paper firstly analyzed the impact of higher or lower pricing, and then investigated and summarized the posted pricing methods for crowdsourcing proposed so far. Secondly, aiming at solving the pricing problem for crowdsourcing, this paper innovatively adopted a machine learning method – Random Forests, and made a detailed theoretical description with in-depth analysis of the four parameters that affect the performance of this algorithm. Python spider technology had been used to acquire real crowdsourcing transaction data. In the final experiment, the actual effects of 8 ML methods in price prediction had been compared, and it can be confirmed that RF almost has the best pricing possibility. In order to adjust and optimize the RF model, two additional experimental schemes is designed, and a comparison between the algorithm's pricing results and training times for modeling with four parameters is made. Finally, by in-depth analysis of the experimental results, a parameter setting suitable for the case in this paper is given. What's more, although it is feasible to solve posted pricing problems with RF, the posted method's results are slightly unsatisfactory. In the future, we will try to regard posted pricing as a classification problem and hope it can perform well in the accuracy of predicting recommended price range.

Acknowledgement. This work was supported by a grant from the National Natural Science Foundation of China (No. 41761086).

References

1. Taskcn (2006). http://www.taskcn.com/
2. Breiman, L., Cutler, A.: Random forests (1999). https://www.stat.berkeley.edu/%7Ebreiman/RandomForests/cc_home.htm
3. Chunmei, Z., Guomei, H., Zuojie, P.: A study of web information extraction technology based on beautiful soup. J. Comput. **10**(6), 381–387 (2015). https://doi.org/10.17706/jcp.10.6.381-387

4. Han, K., Huang, H., Luo, J.: Quality-aware pricing for mobile crowdsensing. IEEE/ACM Trans. Networking **26**(4), 1728–1741 (2018). https://doi.org/10.1109/TNET.2018.2846569
5. Han, K., Huang, H., Luo, J.: Posted pricing for robust crowdsensing. In: ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2016), Paderborn, Germany, pp. 261–270, July 2016. https://doi.org/10.1145/2942358.2942385
6. Holmes, G., Donkin, A., Witten, I.H.: WEKA: a machine learning workbench. In: Proceedings of ANZIIS 1994 - Australian New Zealand Intelligent Information Systems Conference, pp. 357–361, November 1994. https://doi.org/10.1109/ANZIIS.1994.396988
7. Howe, J.: Crowdsourcing: why the power of the crowd is driving the future of business. J. Consum. Mark. **26**(4), 305–306 (2009). https://doi.org/10.1108/07363760910965918
8. Hu, Z., Zhang, J.: Optimal posted-price mechanism in microtask crowdsourcing. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, pp. 228–234 (2017). https://doi.org/10.24963/ijcai.2017/33
9. Kantardzic, M.: Data-Mining Concepts, Chap. 1, pp. 1–25. Wiley, Hoboken (2011). https://doi.org/10.1002/9781118029145.ch1
10. Sun, J., Ma, H.: Collection-behavior based multi-parameter posted pricing mechanism for crowd sensing. In: 2014 IEEE International Conference on Communications (ICC), pp. 227–232, June 2014. https://doi.org/10.1109/ICC.2014.6883323
11. Tariq, S.: Developing market sentiment indicators for commodity price forecasting using machine learning. Master thesis, The University of Manitoba (2018)
12. Xiaojuan, Z.: Application of crawler technology based on python. Office Informatization (2018)
13. Yan, J., Liu, R., Liu, H.: A literature review of domestic and foreign crowdsourcing research. Forum Sci. Technol. China **1**(8), 59–68, 151 (2017)