Crime prediction: pattern recognition and prediction of collusive bidding

Yiming Zhai¹, ², Junqing Cheng², Ning Ding¹, ², *, Zhenyu Fan¹, ², Yang Lu²

¹Public Security Behavioral Science Lab, People’s Public Security University of China, Beijing, China
²College of Investigation, People’s Public Security University of China, Beijing, China
Email: dingning@ppsuc.edu.cn

Abstract. Bidding is an effective means to promote fair competition in the market and optimize the allocation of social resources. However, with the increasing incidence of bid-rigging crimes, public resource regulatory agencies and public security agencies are facing new challenges. On the premise of fully studying the current trend of China’s domestic collusive bidding crimes, this paper summarized the abnormal characteristics of four categories of companies suspected of bid rigging, and selected seven representative subcategories of data based on actual conditions, and set thresholds for these data. Then machine learning methods were utilized to build a pre-warning model for collusion bidding. The results show that the use of machine learning methods to predict the bid collusion can achieve a maximum accuracy of 95%. The research can not only improve the early warning, monitoring, analysis and judgment capabilities of the economic crime investigation department, but ensure the benign operation of bidding, and protect the sustained and rapid development of the social economy.

1. Introduction

At present, as the country’s rigid demand for economic development, bidding, represented by infrastructure construction such as housing construction and highway engineering, is extremely active, and the scale of investment continues to expand. Driven by huge economic benefits, collusive bidding crimes in these areas have also intensified[1].

The crime has intensified due to many reasons, such as the lack of strict qualification review of bidders, the faultiness of supervision mechanism, and the imperfections of the legal and regulatory system. Coupled with the rampant phenomenon of the affiliated contract business and borrowing qualifications behavior in the construction market, this type of crime presents the characteristics of concealment of the process, difficulty of being discovered, diversification of criminal subjects, and collectivization. In particular, the issue of "internal and external collusion" among multiple departments and individuals, such as bid participants and bid evaluation experts, should arouse great attention. In order to change this situation, it is found that in the entire process of implementing collusive bidding, abnormal data will not be erased. Through in-depth analysis of certain abnormal data characteristics in collusive bidding, data rules can be discovered, and data models can also be constructed based on machine learning algorithms.

This paper analyzes the behavioral pattern of the crime of collusive bidding, extracts the main characteristic data, and predicts crime by using different machine learning algorithms. The part I of the
article is the introduction; literature review is set in part II; part III introduces the data source and methods; results and discussions are presented in part IV.

2. Literature review
In terms of legislation on the crime of collusive bidding. He Hongfeng believes that Article 223 of the Criminal Law is relatively simple, and it appears to be somewhat stretched to deal with the complex and diverse bid-rigging situations in the practice of bidding. It is inconsistent with the subsequent "Bidding Law", "Government Procurement Law" and "Regulations for the Implementation of the Bidding Law" in the understanding of related issues, leading to difficulties in judicial application[2]. Based on the core-peripheral social network analysis, Reeves-Latour & Morselli investigated the illegal bidding indicators in more than 7,000 public construction bidding project data in Laval City (Canada) from 1966 to 2013. The results show that companies suspected of manipulating bidding activities are long-term core participants. This is mainly due to the fact that the companies involved in the case that have close ties with the government have become the dominant force in the field of construction engineering and can seek the greatest profit under the conditions of compliance with laws and regulations[3]. Jaspers believes that the legislative policy of lenient penalties provides a good opportunity for companies suspected of bid collusion to confess, and it also provides powerful investigation protection for the investigation department. Many companies involved in the case will choose to surrender themselves in order to avoid criminal penalties or reduce financial fines.

In the investigation of collusive bidding crime. Yang pointed out that when investigating such cases, the economic investigation department of the public security organ should choose the correct investigation methods and investigation channels, obtain evidence in a comprehensive manner, so as to better combat the crime of bid collusion and escort the market economy[4]. Wang emphasized that economic crime investigation departments should take the initiative to adapt to the big data investigation model, and the collusive bidding investigation model should be transformed into "data-based"[5]. Kim et al. proposed a method to estimate the damage caused by bid collusion in a design-build (DB) construction project. It uses a simulation model based on the historical bidding data of each bidder. The model can reproduce the competitive bidding process in the database. The feasibility of the model is proved by case studies of actual railway construction projects. This suggestion model can also be used to determine the best bidding price for DB construction projects and provide ideas for the investigation of collusive bidding cases[6].

3. Data collection and methodology

3.1. Data of collusive bidding
Based on the past experience in handling bid collusion cases, this paper first investigated the Public Resource Exchange Center and the Public Resource Exchange Supervision and Administration Bureau. Then we selected the following four categories of most representative characteristic data: key indicator data, abnormal behavior data, abnormal process data and other abnormal data.

3.1.1. Key indicator data
**Document production machine code (DPMC).** DPMC refers to the identification code formed by MD5 encryption of the five characteristic codes of the computer's MAC address, hard disk number, motherboard number, CPU number, and tool identification number. DPMC can be used as a basis for judging whether bidders use the same computer to make bids.

**Document creation identification code (DCID).** DCID means that when a new bidding document is created in the special production software of the Public Resource Trading Centre, a string of identification numbers will be automatically generated. The identification number is unique and will not change with the modification of the content. DCID can be used to determine whether the bidder uses the same bidding document for modification and upload.
The lock number of the bill of quantities (LNBQ). LNBQ refers to the unique identifier that the engineering project bidder inserts the list encryption lock when editing the bill of quantities. LNBQ can be used as a basis for bidding participants to use the same list encryption lock to make a list.

The bidding document upload IP (BUIP). BUIP refers to the network environment identifier during the bidding document upload process. It can be used as a basis for bidders to upload bidding document in the same network environment.

The remotely decrypted IP (RDIP). RDIP refers to the network identification of the bidder when the bid is opened remotely. It can be used as the basis for remote decryption by bidders in the same network environment.

Document download IP (DDIP). DDIP refers to the network identification when the bidder downloads the bidding documents. It can be used as a basis for bidders to download bidding documents in the same network environment.

3.1.2. Abnormal behavior data

Abnormal bidding behavior means that the winning rate is too high or too low in all projects in which the bidder participates. There may be the following four abnormal behaviors.

1) Bidding collusion is generally a group behavior. The downloading, production and uploading of bidding documents is done by a dedicated person, who usually only downloads one bidding document. Therefore, there will be an abnormal situation where the bidder uploads the bid but has not downloaded the bidding document.

2) According to the on-site situation, the bidder actively contacts the accompanying bidder to give up entering the bid opening, so as to control the bid base price and increase the probability of winning the bid. Therefore, there will be abnormal phenomena that cannot participate in the bid opening process, such as the bid documents are returned or not decrypted.

3) The bidding documents are tailor-made for the default winning bidder in the bidding documents, which caused other bidders to abandon the bidding because they could not meet the requirements after purchasing the bidding documents.

4) The collusive bidding group will organize unified uploading of bids, and the uploading time will be periodic. Through the analysis of the time aggregation of the uploading, the bidders with abnormal behavior can be found.

3.1.3. Abnormal process data

The abnormal bidding process has the following situations:

1) The complaint is abnormal. Projects that are not rigorous in the content of the bidding documents or have violations of laws and regulations during the bidding process and are questioned or complained by bidders are likely to have a risk of colluding bidding.

2) The number of bidders is abnormal. The contract-issuing party tailor-made bidding documents for specific objects, causing other bidders to abandon their bids[7]. Or some of the bidders involved gang-related companies, which prevented competitive companies from participating in bidding through special means.

3) The bidding failed. The contract-issuing party raised the threshold in the bidding documents, resulting in the bidders not reaching the legal number and re-tendering, so as to achieve the purpose of direct entrustment after twice.

4) Abnormal rate of rejection of bids. The bid was invalidated because the bidder accompanies the bid casually, or because the bidding party deliberately increased the bid threshold, the number of invalid bids was too large.

5) Postponement of the bidding process. An accident occurred during the bidding process, and the project that applied for the extension of the bidding and the time exceeded the prescribed time limit may be caused by the failure to allocate the interests of the parties involved in the collusive bidding[8].
3.1.4. Other abnormal data
Other abnormal data include:

1) Abnormal bid price. Under the premise of ensuring the winning of the bid, the collusive bidding group will ask some bidders to quote high prices and raise the average price in order to pursue profits, thereby increasing the winning bid amount.

2) The biased evaluation of experts. By analysing the scoring data of the bid evaluation projects in which the experts participate, the experts may give high scores due to interest-related tendencies.

3) Identical errors in tenders. Through the analysis of all bidding documents under the same project, bidders with similar errors may be suspected of colluding bidding.

4) Abnormal bidding contact. The bidder will register the bidding contact person on the trading platform. The bidding contact person generally belongs to only one bidding unit. If it appears in different bidding units, it largely reflects the close relationship between the units.

3.2. Application of machine learning algorithms
To establish prediction models for the crime of bid collusion, it is necessary to extract the abnormal behavior elements of the company suspected of bid collusion in the entire bidding process. At the same time, it is necessary to consider that due to the confidentiality of some cases, the characteristic data should not be retrieved or disclosed. Therefore, this paper selected 7 types of characteristic data of companies suspected of colluding with bidding, including upload interval (minutes), joint bidding, bidding contact, MAC address, IP address, bidding behavior, and low bid winning rate.

Based on the experience of the police and the competent administrative agency in handling such cases, different thresholds have been set for the 7 categories of data: the interval between the bidding documents uploaded by the accompanying bidder and the winning bidder is less than 60 minutes; the joint bid rate is greater than 30%; the contact person of different bidding companies is same; mac address of the bidding documents (DPMC, DCID) is same; BUIP, RDIP and DDIP are same; behaviors such as bid questioning and invalidation in the process; the winning rate is lower than 2%.

In order to test the rationality of the above threshold setting, the 120 data collected from the road appearance improvement project in the economic and technological development zone of a certain city in Anhui Province were analysed. The data included 60 colluding bidding companies and 60 normal bidding companies. In these projects, companies that have been punished by administrative agencies, interviewed, or investigated by public security agencies are deemed to be suspected of colluding in bidding.

| Categories                        | Collusive bidding | Normal bidding |
|-----------------------------------|-------------------|----------------|
| Time interval for uploading bids (minutes) | 93.33%           | 0.00%          |
| Joint bidding rate                | 96.67%           | 0.00%          |
| Same bid contact                  | 50.00%           | 8.33%          |
| MAC address                       | 61.67%           | 15.00%         |
| IP address                        | 95.00%           | 36.67%         |
| Low bid winning rate              | 95.00%           | 10.00%         |
| Bidder not participating in the bid opening | 35.00%           | 5.00%          |

As shown in Table 1, among the companies suspected of bid collusion, 93.33% of the companies’ bid upload interval is less than 60 minutes, 96.67% of the companies’ joint bid rate is greater than 30%, and half of the companies have the same contact person for the bid, 61.67% of the companies’ MAC addresses are abnormal, 95.00% of the companies’ IP addresses are abnormal, 35.00% of the companies’ bidding behaviors are abnormal, and 95.00% of the companies’ bid winning rate is abnormally low.

Among the normal bidding companies, there is no company whose upload interval is less than 60 minutes, and the joint bidding rate is less than 30%. 8.33% of the companies have the same contact
person, 15.00% of the companies have abnormal mac addresses, 36.67% of the company’s IP address is abnormal, 10.00% of the companies had abnormal bidding behavior, and 5.00% of the companies had an abnormally low bid winning rate. It can be seen that by setting the threshold, the distinction between colluding bidders and normal bidders is more obvious.

Now more and more scholars use machine learning methods to analyse and predict crimes, and have achieved better results. For example, machine learning methods are used to analyse and predict bus pickpocketing crimes[9], credit card fraud crimes[10], and fraudulent invoicing crimes[11]. In this article, in order to test which machine learning model is more suitable for analysing the crime of collusive bidding, the data are imported into the support vector machine, naive Bayes, logistic regression and random forest algorithms in Scikit-learn. The company suspected of colluding in bidding is represented by the label "+1", and the normal company is represented by the label ".-1". With 70% of the data as the training set and 30% of the data as the test set. The accuracy of classification is shown in Table 2.

4. Results and discussions

It can be seen from the results that all five algorithms have achieved good classification results, and their accuracy rates are all above 84%, of which the highest can reach 95%. Therefore, it is feasible to use machine learning methods to monitor and warn the crime of collusive bidding. But the prerequisite is to have rich experience and the ability to accurately identify the key elements of the case. Thus, the data can be preprocessed to a certain extent, which is the key to improving the accuracy of machine learning.

| Algorithm     | Category | Precision | Recall | F1-Score | Support |
|---------------|----------|-----------|--------|----------|---------|
| SVM           | 0        | 0.91      | 0.83   | 0.87     | 24      |
|               | 1        | 0.71      | 0.83   | 0.77     | 12      |
|               | weighted avg | 0.84   | 0.83   | 0.84     | 36      |
| Decision Tree | 0        | 0.89      | 1      | 0.94     | 24      |
|               | 1        | 1         | 0.75   | 0.86     | 12      |
|               | weighted avg | 0.93   | 0.92   | 0.91     | 36      |
| Random Forest | 0        | 0.92      | 1      | 0.96     | 24      |
|               | 1        | 1         | 0.83   | 0.91     | 12      |
|               | weighted avg | 0.95   | 0.94   | 0.94     | 36      |
| Logistic Regression | 0        | 0.92      | 1      | 0.96     | 24      |
|               | 1        | 1         | 0.83   | 0.91     | 12      |
|               | weighted avg | 0.95   | 0.94   | 0.94     | 36      |
| Naïve Bayes   | 0        | 0.92      | 1      | 0.96     | 24      |
|               | 1        | 1         | 0.83   | 0.91     | 12      |
|               | weighted avg | 0.95   | 0.94   | 0.94     | 36      |

Through timely retrieval of back-end data, and the use of machine learning methods to conduct real-time monitoring and early warning of companies suspected of bid collusion, the current problem of unable to find clues in time can be effectively solved. The root of these difficulties lies in the lack of data. Therefore, taking big data as the starting point and using big data penetrating means to develop collusive bidding warning models has become an effective way to solve this problem. The better modeling idea is to comprehensively collect the underlying data in the bidding process and actively discover clues.

The method in this article still has certain limitations, which are mainly reflected in: 1) the types of data obtained are not complete, and some key data are difficult to obtain; 2) the number of companies involved in the case is small, and more cases of suspected bid-rigging crimes need to be tested; 3) the
police and the bidding administrative department respectively have sensitive data, and it is difficult to integrate the data.

References
[1] H. Yuan, "Discussion and Analysis on Some Problems in the Bidding and Bidding of Construction Projects and Their Solutions," Technology Innovation and Application, p. 229-229, 2014.
[2] H. He, "Modification of Correlation Laws," Journal of Tianjin Administrative Institute of Political Science & Law, pp. 5-10, 2012.
[3] M. Reeves-Latour and C. Morselli, "Bid-rigging networks and state-corporate crime in the construction industry," Social Networks, p. S0378873316302386, 2016.
[4] Y. Yi-kun, "The Study of the Harm, Causes and Countermeasures of the Collusive Tendering Cases," Journal of Qiqihar University (Philosophy and Social Science Edition), pp. 89-91, 2017.
[5] C. Wang, "On the Investigation Difficulties and Solutions of Bidding Collusive Cases——Based on the investigation of the "three supply and one industry" mega-bidding collusion case in F City, L Province," Journal of Jiangxi Public Security College, pp. 33-40, 2020.
[6] J. K. Kyong, K. Kyoungmin, W. K. Eu, and H. T. N. Thi, "Estimating Damages of Bid-Rigging in Design-Build Contracts Based on Simulation Model," KSCE Journal of Civil Engineering, 2021-02-19 2021.
[7] Z. Cao, "STUDY ON RISK MANAGEMENT AND CONTROL OF BIDDING AND TENDERING IN UNIVERSITY,". vol. Master: Southwest Jiaotong University, 2015.
[8] L. Luo,"Study on the Legal Regulation relating to Bid——rigging in Government Procurement,". vol. Master: Zhengzhou University, 2013.
[9] N. Ding and Y. Zhai, "Crime prevention of bus pickpocketing in Beijing, China: does air quality affect crime?" Security Journal, 2019.
[10] C. Li, N. Ding, Y. Zhai, and H. Dong, "Comparative study on credit card fraud detection based on different support vector machines," Intelligent Data Analysis, vol. 25, 2021-01-26 2021.
[11] D. Ning, Z. Xinnan, Z. Yiming, and L. Chenglong, "Risk assessment of VAT invoice crime levels of companies based on DFPSVM: a case study in China," Risk Management, 2021-03-13 2021.