



# Role Recognition of Illegal Online Gambling Participants Using Monetary Transaction Data

Xiaohui Han<sup>(✉)</sup>, Lianhai Wang, Shujiang Xu, Dawei Zhao, and Guangqi Liu

Shandong Computer Science Center (National Supercomputer Center in Jinan),  
Shandong Provincial Key Laboratory of Computer Networks, Qilu University of  
Technology (Shandong Academy of Sciences), Jinan, China  
{hanxh,wanglh,xushj,zhaodw,liuguangqi}@sdas.org

**Abstract.** Online gambling has become a substantial global industry during the past two decades. However, it is explicitly prohibited or restricted by most countries in the world due to social problems caused by it. This results in rapid expansion of the illegal online gambling (IOG) market where players profits are under little protection. To fight against IOG, this paper addresses the IOG participant-role recognition (PRR) problem by learning a supervised classifier with monetary transaction data. We propose two sets of features, i.e., transaction statistical features and network structural features, to effectively represent participants. Based on the feature representation, we adopt an ensemble learning strategy in the training phase of a PRR classifier to reduce the impact of unbalanced data. Results of experiments performed on real-world IOG case data demonstrate the feasibility and validity of the proposed approach. The proposed approach could help investigators in a law enforcement agency find the key members of an IOG organization quickly and destroy the ecosystem efficiently.

**Keywords:** Illegal online gambling · Role recognition · Online crime  
Monetary transaction

## 1 Introduction

Online gambling, which was initiated in the mid-1990s, has exploded from a minor sideshow on the Internet into a substantial global industry over the past two decades. Due to factors such as convenience, accessibility, affordability, anonymity, and interactivity, online gambling could be potentially more tempting and addictive to consumers than traditional offline gambling. However, several recent studies suggest that it would exacerbate gambling problems in society, such as problem gambling, pathological gambling, and underage gambling [6, 8]. Consequently, a number of countries, including Mainland China, the United States, and Russia, explicitly prohibit most or all forms of online gambling. In the most remaining countries, online gambling is under strict regulations.

The prohibition and restriction of online gambling result in rapid expansion of the illegal online gambling (IOG) market where players' profits are under little protection. Several studies allege that IOG sites victimize participants rather than benefit them [1–4]. More seriously, it is reported that the IOG business is correlated with money laundering and other complex fraudulent and extortionist activities that, in turn, can fund yet other criminal activities [3]. Although states have adopted legal measures targeted at both consumers and operators to limit citizens' access to IOG sites. Yet the effectiveness of such measures may well be limited.

To fight against IOG, it is necessary to gain sufficient knowledge regarding how these businesses operate in an environment characterized by extreme uncertainty and high risk. This would require a systematic analysis of sophisticated organizational structures formed by IOG participants. However, to the best of our knowledge, fairly few researches have been carried out on automatic techniques for analyzing IOG ecosystems. This study addresses this absence by investigating the problem of automatically recognizing the roles that IOG participants played in their ecosystem. The proposed approach could help investigators in a law enforcement agency (LEA) find the key members of an IOG organization quickly and destroy the ecosystem efficiently. To be specific, this study makes the following contributions:

- We present a novel participant-role recognition (PRR) approach which learns a supervised classifier based on monetary transaction data to predict the roles of IOG participants.
- We propose two sets of features, i.e. transaction statistical features and network structural features, to effectively represent participants.
- We adopt an ensemble learning strategy in the training phase of the PRR classifier to reduce the impact of unbalanced training data.
- We evaluate the performance of the proposed approach using real-world IOG monetary transaction data. Experimental results demonstrate the feasibility and validity of the proposed approach.

The remainder of this paper is organized as follows: Sect. 2 reviews related work. Section 3 provides preliminaries related to this paper. Section 4 describes the details of the proposed approach. Section 5 gives the experimental results and analysis. Section 6 concludes the content of this paper.

## 2 Related Work

Prosperity on Internet gambling has drawn much attention from academics [11]. A number of studies have investigated possible precipitating factors for the expansion of online gambling. For example, Gainsbury et al. [6] claimed that technological innovations, including the availability of cheap, fast broadband connections in essentially any location, the emergence of mobile technology, and the use of trustable online payment systems, have played an important role in



**Fig. 1.** The pyramid structure of an IOG organization.

the growth of online gambling. They also found that the primary reasons people gave for preferring to gambling were: ease of access, convenience, comfort, greater privacy, and anonymity. Similar motivations for online gambling have been found in [14]. Several pieces of research studied demographic profiles of online gamblers and found that Internet gamblers were significantly more likely to be male gender, younger age, from higher socio-economic strata, employed full time, more technologically savvy, having more positive attitudes toward gambling, and better educated [6, 8, 9, 12, 13]. However, it was reported more women and young people are engaging in the activity [7].

Many recent studies reported online gambling is more harmful to gamblers compared to terrestrial gambling [6, 8, 9]. Wood and William [13] found online gamblers are under a higher probability of using drugs and alcohol than non-online gamblers. Due to the fragmented nature of governance, online gambling also presents a significant opportunity for crime and victimization. McMullan and Rege [10] investigated the types, techniques, and organizational dynamics of crime at portals of online gambling sites using document analysis based on data retrieved with the Google search engine.

In jurisdictions, such as the United States and Mainland China, for example, the prohibition of gambling has given rise to illegal online gambling business. Banks [3] summarized three principal forms of the illegal gambling business, i.e., accepting bets from a resident in a country where gambling is illegal, operating without appropriate licenses, and accepting bets from underage gamblers. There are numerous documented cases in which providers of illegal online gambling have been found to cheat customers on payouts, have apparently not paid winnings, have cheated players with unfair games, or have absconded with player deposits [1–4]. As claimed by McMullan and Rege [10], these cases identified by existing studies are likely to represent “the tip of the iceberg”, with many more crimes going unreported and unrecorded.

Based on our review of the related work, we have found that although IOG has been studied for years, most existing researches only focused on types of crime related to IOG. There continues to be a paucity of research on the systematic analysis of sophisticated organizational structures IOG ecosystems.

Especially, few automatic techniques have been proposed to analyze IOG ecosystems. In this study, we address this absence by presenting an approach to automatically recognize the roles that IOG participants played in their ecosystem.

### 3 Preliminaries

#### 3.1 The IOG Ecosystem and Participant Roles

Figure 1 shows the typical structure of an IOG ecosystem, which is like a pyramid. A small number of **investors** provide funds for building online gambling platforms in countries/regions where online gambling is legal. Then these investors seek local organizers in a country/region, where online gambling is forbidden, to run the IOG business in that country/region. **Local organizers** employ gambling agents to attract gamblers to participate in IOG games. **Gambling agents** collect bets from gamblers and distribute IOG proceeds among participants. **Gamblers** can only participate in IOG games by transferring their bets to gambling agents. Participant roles in a higher level of the pyramid structure will keep interest at a definite ratio from the proceeds they collected from the direct lower level role. Based on the understanding of the IOG ecosystem, we aim to identify IOG participants into one of the four roles, i.e., investor, local organizer, gambling agent, and gambler.

#### 3.2 Problem Statement

The participant-role recognition (PRR) problem can be considered as a supervised classification problem. We give a formal definition of PRR as follows.

**Definition 1.** *Given a set of  $K$  pre-defined roles  $\mathcal{R} = \{r_1, r_2, \dots, r_K\}$  and a set of IOG participants  $\mathcal{P} = \{p_1, p_2, \dots, p_M\}$ , the task of participant-role recognition is to find a function  $f(\cdot)$  to decide if a participant  $p_i$ 's role is  $r_j$ , i.e.,  $f : \mathcal{P} \times \mathcal{R} \mapsto \{0, 1\}$  such that for any pair  $(p_i, r_j) \in \mathcal{P} \times \mathcal{R}$ , we have*

$$f(p_i, r_j) = \begin{cases} 1, & \text{if } p_i \text{'s role is } r_j \\ 0, & \text{otherwise} \end{cases}. \quad (1)$$

For the  $i$ -th participant, we represent them as a  $n$ -dimensional feature vector  $p_i = [p_{i1}, p_{i2}, \dots, p_{in}]^T$ , where  $p_i \in \mathbb{R}^n$ . By this definition, solving the PRR problem involves extracting discriminative features as well as finding an appropriate classification scheme. We present our solution in the following section.

### 4 The Proposed Approach

Figure 2 provides an illustration of our PRR approach. We first compute transaction statistics and build a money flow network based on monetary transaction data obtained from an IOG ecosystem. Then we extract transaction statistical features and network structural features to form a vector representation for each participant sample. Based on the extracted features, we further learn a classifier to predict the role of new participant samples. This section describes each component of the approach in detail.

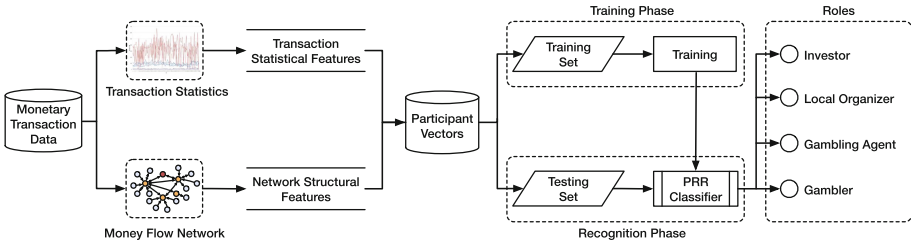


Fig. 2. Illustration of the proposed participant-role recognition approach.

### 4.1 Monetary Transaction Data

Monetary transaction data (MTD), which records the monetary transactions between IOG participants, could be obtained from bank accounts of IOG participants or derived from IOG servers. Each transaction record typically consists of information including the sender, the recipient, the amount of money transferred, and a time stamp. MTD contains very useful information for understating an IOG ecosystem. However, it is usually difficult to access MTD due to privacy reasons. The MTD used in this study was provided by a law enforcement agency. They obtained the data with warrants during the investigation of an IOG case.

### 4.2 Feature Extraction

We propose two sets of features to represent IOG participants in the learning process of a PRR classifier.

**Transaction Statistical Features.** We assume that participants in different roles have discriminative money transfer patterns which can be captured by statistics extracted from MTD. Here, we first introduce some notations used in the computation of transaction statistical features. Let  $M$  be the number of participants in an IOG ecosystem, and let  $N_D$  be the number of days covered by the MTD. We denote  $\mathbf{A}^{(in)}$ ,  $\mathbf{A}^{(out)}$ ,  $\mathbf{F}^{(in)}$ ,  $\mathbf{F}^{(out)}$ ,  $\mathbf{C}^{(in)}$ , and  $\mathbf{C}^{(out)}$  as  $M \times N_D$  dimensional matrices, where  $\mathbf{A}_{ij}^{(in)}$  ( $\mathbf{A}_{ij}^{(out)}$ ) is the amount of money participant  $p_i$  received (sent out) on the  $j$ -th day,  $\mathbf{F}_{ij}^{(in)}$  ( $\mathbf{F}_{ij}^{(out)}$ ) is the number of transfers  $p_i$  received (sent out) on the  $j$ -th day, and  $\mathbf{C}_{ij}^{(in)}$  ( $\mathbf{C}_{ij}^{(out)}$ ) is the number of incoming (outgoing) counterparties of  $p_i$  on the  $j$ -th day, respectively. To capture money transfer patterns of IOG participants, we compute the following features.

- **Means of Daily Income ( $\hat{\mu}_i^{(ain)}$ ) and Expenditure ( $\hat{\mu}_i^{(aout)}$ )** measure the average amounts of money participant  $p_i$  receives from and sends to other participants daily, which are computed by the following equations:

$$\hat{\mu}_i^{(ain)} = \frac{1}{N_D} \sum_{j=1}^{N_D} \mathbf{A}_{ij}^{(in)}, \tag{2}$$

$$\hat{\mu}_i^{(aout)} = \frac{1}{N_D} \sum_{j=1}^{N_D} \mathbf{A}_{ij}^{(out)}. \quad (3)$$

- **Variances of Daily Income** ( $\hat{\sigma}_i^{(ain)}$ ) **and Expenditure** ( $\hat{\sigma}_i^{(aout)}$ ) measure how far the daily incomes and expenditures of participant  $p_i$  are spread out from the mean values, which are computed by the following equations:

$$\hat{\sigma}_i^{(ain)} = \frac{1}{N_D - 1} \sum_{j=1}^{N_D} [\mathbf{A}_{ij}^{(in)} - \hat{\mu}_i^{(ain)}]^2, \quad (4)$$

$$\hat{\sigma}_i^{(aout)} = \frac{1}{N_D - 1} \sum_{j=1}^{N_D} [\mathbf{A}_{ij}^{(out)} - \hat{\mu}_i^{(aout)}]^2. \quad (5)$$

- **Means of Daily Incoming** ( $\hat{\mu}_i^{(fin)}$ ) **and Outgoing Transfer Numbers** ( $\hat{\mu}_i^{(fout)}$ ) measure the average numbers of daily transfers participant  $p_i$  receives and sends out, which are computed by the following equations:

$$\hat{\mu}_i^{(fin)} = \frac{1}{N_D} \sum_{j=1}^{N_D} \mathbf{F}_{ij}^{(in)}, \quad (6)$$

$$\hat{\mu}_i^{(fout)} = \frac{1}{N_D} \sum_{j=1}^{N_D} \mathbf{F}_{ij}^{(out)}. \quad (7)$$

- **Variances of Daily Incoming** ( $\hat{\sigma}_i^{(fin)}$ ) **and Outgoing Transfer Numbers** ( $\hat{\sigma}_i^{(fout)}$ ) measure how far  $p_i$ 's daily numbers of incoming and outgoing transfers are spread out from their mean values, which are computed by the following equations:

$$\hat{\sigma}_i^{(fin)} = \frac{1}{N_D - 1} \sum_{j=1}^{N_D} [\mathbf{F}_{ij}^{(in)} - \hat{\mu}_i^{(fin)}]^2, \quad (8)$$

$$\hat{\sigma}_i^{(fout)} = \frac{1}{N_D - 1} \sum_{j=1}^{N_D} [\mathbf{F}_{ij}^{(out)} - \hat{\mu}_i^{(fout)}]^2. \quad (9)$$

- **Means of Daily Incoming** ( $\hat{\mu}_i^{(cin)}$ ) **and Outgoing Counterparty Numbers** ( $\hat{\mu}_i^{(cout)}$ ) measure  $p_i$ 's average numbers of daily incoming and outgoing counterparties, which are computed by the following equations:

$$\hat{\mu}_i^{(cin)} = \frac{1}{N_D} \sum_{j=1}^{N_D} \mathbf{C}_{ij}^{(in)}, \quad (10)$$

$$\hat{\mu}_i^{(cout)} = \frac{1}{N_D} \sum_{j=1}^{N_D} \mathbf{C}_{ij}^{(out)}. \quad (11)$$

– **Variiances of Daily Incoming ( $\hat{\sigma}_i^{(cin)}$ ) and Outgoing Counterparty Numbers ( $\hat{\sigma}_i^{(cout)}$ )** measure how far  $p_i$ 's daily numbers of incoming and outgoing counterparties are spread out from the mean values, which are computed by the following equations:

$$\hat{\sigma}_i^{(cin)} = \frac{1}{N_D - 1} \sum_{j=1}^{N_D} [C_{ij}^{(in)} - \hat{\mu}_i^{(cin)}]^2, \tag{12}$$

$$\hat{\sigma}_i^{(cout)} = \frac{1}{N_D - 1} \sum_{j=1}^{N_D} [C_{ij}^{(out)} - \hat{\mu}_i^{(cout)}]^2. \tag{13}$$

**Network Structural Features.** All IOG participants are in pursuit of money. Gamblers want to win money by playing various kinds of games provided by an IOG platform. Investors, local organizers, and gambling agents aim to earn interest from gamblers. Therefore, money flow can be seen as the “blood” of an IOG ecosystem. We use a money flow network (MFN) to describe the “flow” of “blood”. Let  $f_{ij}$  be the total amount of money transferred from participant  $p_i$  to participant  $p_j$ , and let  $f_{ji}$  be the total amount of money transferred from  $p_j$  to  $p_i$ . We first define the money flow between  $p_i$  and  $p_j$  as follows.

**Definition 2.** *The money flow between two participants  $p_i$  and  $p_j$  is the absolute value of the difference between  $f_{ij}$  and  $f_{ji}$ , which is computed by:*

$$flow(p_i, p_j) = |f_{ij} - f_{ji}|. \tag{14}$$

*The direction of the flow is from  $p_i$  to  $p_j$  if  $f_{ij} > f_{ji}$ , and is from  $p_j$  to  $p_i$  if  $f_{ji} > f_{ij}$ .*

Based on the definition of money flow, we further give the definition of an MFN.

**Definition 3.** *A money flow network is a directed graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V}$  is a finite set of vertices representing participants, and  $\mathcal{E}$  is a set of directed edges representing money flows between participants.*

In  $\mathcal{G}$ , there will be an edge  $e_{ij} \in \mathcal{E}$  between two vertices  $v_i$  and  $v_j$  in  $\mathcal{V}$  if  $flow(p_i, p_j) \neq 0$ , and the weight and direction of the edge are the same with  $flow(p_i, p_j)$  and the direction of the flow, respectively. An MFN can further be represented by an  $M \times M$  dimensional adjacency matrix  $\mathbf{W}$ , where

$$\mathbf{W}_{ij} = \begin{cases} flow(p_i, p_j), & \text{if } e_{ij} \in \mathcal{E} \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

We assume that the importance of different roles in an MFN should be different. Thus, we compute the following centrality measures as network structural features to identify the importance of participants in an MFN.

- **In-degree.** This measure counts the number of incoming ties of a vertex. As edges in an MFN are weighted, values of incoming edge weights are summed up. The in-degree for vertex  $v_i$  is represented by the equation:

$$D_{in}(v_i) = \sum_{j=1}^M \mathbf{W}_{ji} \tag{16}$$

- **Out-degree.** This measure counts the number of out-going ties of a vertex. The weighted out-degree for vertex  $v_i$  is represented by the equation:

$$D_{out}(v_i) = \sum_{j=1}^M \mathbf{W}_{ij} \tag{17}$$

- **All-degree.** This measure counts the number of ties connecting a vertex to the others, regardless of their directionality. In the weighted version, the all-degree for vertex  $v_i$  is represented by the equation:

$$D_{all}(v_i) = \sum_{j=1}^M (\mathbf{W}_{ij} + \mathbf{W}_{ji}) \tag{18}$$

- **Betweenness.** Degrees are local measures, i.e., they do not take into account the whole network, but only the local neighborhood of a vertex. As a global measure, betweenness quantifies how frequently a vertex acts as a bridge along the shortest paths that connect every other couple of vertices. More formally, the betweenness centrality of node  $v_k$  is computed by the following equation:

$$C_b(v_k) = \frac{1}{(M-1)(M-2)} \sum_{i,j \neq k} \frac{N_{ij}^s(v_k)}{N_{ij}^s}, \tag{19}$$

where  $N_{ij}^s$  is the number of shortest paths linking the couple of vertices  $v_i$  and  $v_j$ , and  $N_{ij}^s(v_k)$  is the number of that paths which contain  $v_k$ .  $[(M-1)(M-2)]$  is the total number of pairs of vertices not including  $v_k$ .

- **Closeness.** This feature measures the inverse of the distance of a vertex from all the others in the network, considering the shortest paths that connect each couple of vertices. That is, it denotes how close a vertex is to others. Let  $dist(v_i, v_j)$  be the number of edges in the shortest path linking vertices  $v_i$  and  $v_j$ , the closeness centrality of the vertex  $v_i$  is computed by the following equation:

$$C_c(v_i) = \frac{1}{\sum_{j=1}^M dist(v_i, v_j)}, \tag{20}$$

where  $\sum_{j=1}^M dist(v_i, v_j)$  is the distance of vertex  $v_i$  from all the other vertices in the graph.



### 4.3 Participant Role Recognition

After we determine the features, we use them to represent IOG participants as vectors. Based on the vector representation, we train a classifier using labeled data for role recognition. One thing should be noted is that there may be an imbalance between the number of samples in each role category. For example, in an IOG ecosystem, the number of gamblers is much larger than that of investors. This imbalance could affect the performance of the trained classifier.

To reduce the negative impact of unbalanced training data, we adopt an ensemble learning strategy based on the AdaBoost algorithm [5]. We use Naive Bayesian as the base classifier. Each training tuple is assigned with a weight. A series of  $t$  classifiers are iteratively learned. In each learning iteration, the samples from the original training set are re-sampled to form a new training set. The samples with higher weights are selected with a higher probability. After a classifier  $H_i$  is learned, the samples misclassified by  $H_i$  are assigned higher weights. In the following learning iteration, the classifier  $H_{i+1}$  will pay more attention to the misclassified samples. In each round, we restrict the number of re-sampled samples in each role category to be the same (i.e., the size of the smallest category). The final class prediction is based on the weighted votes of the classifiers learned in each iteration. We named the overall ensemble classifier as EC4PRR (Ensemble Classifier for Participant Role Recognition).

## 5 Experiments and Analysis

In this section, we present a set of experiments to evaluate the performance of our PRR approach using real-world IOG monetary transaction data. We implemented all experiments in Java and MATLAB.

### 5.1 Data

As aforementioned, the MTD used in experiments were provided by an LEA in Shandong Province of China. The data contains two years of monetary transaction records between participants of an IOG platform. The LEA obtained this data during the investigation of the IOG case with warrants, and some of the key members of the IOG organization have been arrested. However, to protect the privacy, the MTD has been pre-processed by the LEA. Each participant was represented by an ID, and only a few fields of each transaction record were kept, including source ID, target ID, amount of transferred money, and time stamp. All the other information was removed. There are totally 4690 participants and 3.9 million transactions in the MTD. Each participant has been assigned a role manually by LEA investigators during the investigation. In Tables 1 and 2, we summarize some essential characteristics of the MTD. From Table 1, we can see the imbalance between the number of samples in each role category. In the training phase, we randomly select 85% of the samples as the training set to build a classifier and 15% as testing samples to validate the performance of the classifier.

**Table 1.** Participants distribution among roles.

	Investor	Local organizer	Gambling agency	Gamblers
# of participants	113	376	1930	2271

**Table 2.** Experimental data statistics.

Statistics	Incoming	Outgoing
Max amount of money a participant transfer daily	998,083 RMB	999,866 RMB
Min amount of money a participant transfer daily	1,000 RMB	1,000 RMB
Max # of daily transfers for a single participant	1136	1419
Min # of daily transfers for a single participant	1	1
Max # of daily counterparties for a single participant	836	655
Min # of daily counterparties for a single participant	1	1

### 5.2 Evaluation Measures

We adopt *precision*, *recall*, and  $F_1$  score as performance measures. Let  $\tilde{\mathcal{R}} = \{\tilde{R}_1, \tilde{R}_2, \dots, \tilde{R}_K\}$  refer to the role assignment output by a PRR classifier, and  $\mathcal{R} = \{R_1, R_2, \dots, R_K\}$  be the ground-truth. Let  $TP_i$  be the number of participants assigned to  $\tilde{R}_i$  correctly according to  $R_i$ ,  $FP_i$  refer to the number of participants assigned to  $\tilde{R}_i$  by mistake, and  $FN_i$  refer to the number of participants should be assigned to  $R_i$  but are assigned to other roles, the *precision*, *recall*, and  $F_1$  for the whole experimental results can be briefed as follows:

$$precision = \frac{\sum_{i \in [1, K]} TP_i}{\sum_{i \in [1, K]} TP_i + \sum_{i \in [1, K]} FP_i} \tag{21}$$

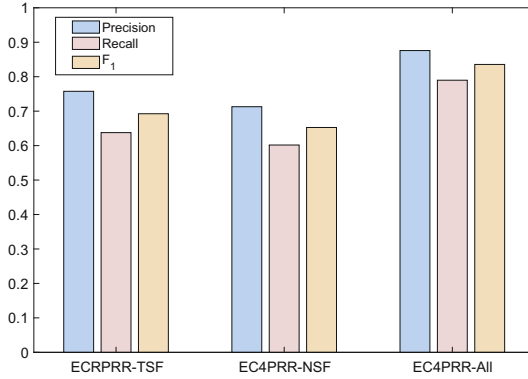
$$recall = \frac{\sum_{i \in [1, K]} TP_i}{\sum_{i \in [1, K]} TP_i + \sum_{i \in [1, K]} FN_i} \tag{22}$$

$$F_1 = \frac{2 \times recall \times precision}{recall + precision} \tag{23}$$

### 5.3 Effectiveness of Features

To tested the effectiveness of the proposed features, we trained EC4PRR with different feature configurations. Three classifiers, named EC4PRR-TSF, EC4PRR-NSF, and EC4PRR-All, were learned using only transaction statistical features, only network structural features, and a combination of all features, respectively. For each classifier, we ran it using a 10-fold cross-validation.

The prediction results are shown in Fig. 3. From the results, we can see both EC4PRR-TSF and EC4PRR-NSF achieved relatively satisfactory performances. Precision scores of EC4PRR-TSF and EC4PRR-NSF achieved 0.76 and 0.71, respectively. The overall performances of the two classifiers are 0.69 and 0.65 respectively in terms of  $F_1$ . This indicates that both transaction statistical



**Fig. 3.** Prediction results of EC4PRR classifiers learned with different feature configurations.

features and network structural features are discriminative for role recognition. EC4PRR-All, which was trained with all features, obtained the best performance in terms of all evaluation measures. The precision, recall, and  $F_1$  of EC4PRR-All were 0.8759, 0.7898, and 0.8356, respectively. This reveals that a combination of the two sets of features can improve the performance of the EC4PRR classifier.

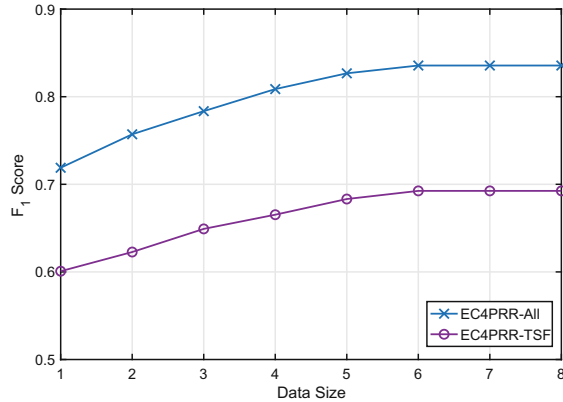
As the efficiency of transaction statistical features can be impacted by the time coverage of MTD, we also examined the sensitivity of EC4PRR-TSF and EC4PRR-All to the size of a dataset. We divided the two years of MTD into eight units with a time coverage of three months for each unit. We then derived subsets of the MTD by varying the time coverage from one to eight units and tested  $F_1$  performances of the two classifiers using these eight subsets. Figure 4 shows the performance variations of EC4PRR-TSF and EC4PRR-All with data size enlarging. We can see that  $F_1$  scores of both EC4PRR-TSF and EC4PRR-All become stable when the data size is larger than five units.

**Table 3.** Performance comparison of different classifiers.

	Naive Bayesian	SVM	Random forest	EC4PRR
Precision	0.5893	0.8286	0.8463	<b>0.8759</b>
Recall	0.5514	0.7832	0.7866	<b>0.7898</b>
$F_1$	0.5697	0.8053	0.8154	<b>0.8356</b>

## 5.4 Role Recognition Results

We compared the proposed EC4PRR with three common-used classifiers, which include Naive Bayesian, SVM, and Random Forest, to validate its predictive



**Fig. 4.** Performance variations of EC4PRR-TSF and EC4PRR-All with the size of data enlarging.

performance. These classifiers have been implemented in WEKA<sup>1</sup> and LIBSVM<sup>2</sup>. We also ran each of the classifiers using a 10-fold cross-validation.

Table 3 gives the results of this experiment. From Table 3, we can see that EC4PRR outperformed all the other three classifiers in terms of precision, recall, and  $F_1$ . This demonstrates the effectiveness of EC4PRR for recognizing the roles of IOG participants. The performance of Naive Bayesian, which we used as the base classifier in our ensemble learning process, was the worst among all classifiers. The ensemble learning strategy made the final classifier much better. Note that the Random Forest is also a kind of classifier based on ensemble learning. However, our modification of the re-sampling strategy made EC4PRR more suitable for the unbalanced training data and achieve better performance.

## 6 Conclusion

In this study, we propose an automatic approach to address the IOG participant-role recognition problem. The proposed approach extracts two sets of features, i.e. transaction statistical features and network structural features, from monetary transaction data to effectively represent participants. A classifier named EC4PRR is trained based on the feature representation to predict the roles of participants. To reduce the impact of unbalanced training data, EC4PRR adopts an ensemble learning strategy in the training phase. Experiments were carried out on real-world IOG case data. The results indicate that both transaction statistical features and network structural features are discriminative for role recognition, and a combination of the two sets of features can improve the performance of EC4PRR. In comparison with other common-used

<sup>1</sup> <http://www.cs.waikato.ac.nz/ml/weka/>.

<sup>2</sup> <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.

classifiers, EC4PRR achieved the best performance. The result also reveals that our modification of the ensemble learning process makes EC4PRR more suitable for unbalanced training data.

There are primarily two limitations of this study we should pay attention to in our future work. First, although we used real-world data in our experiments, the data was only extracted from a specific IOG case, more case data should be collected in the future. Second, we did not consider the situation that an IOG participant plays multiple roles. New techniques should be proposed to address this problem.

**Acknowledgment.** This work was supported in part by National Natural Science Foundation of China (61602281, 61702309), Natural Science Foundation of Shandong Province of China (ZR2015YL018, ZR2016YL011, and ZR2016YL014), Shandong Provincial Key Research and Development Program (2018CXGC0701, 2018GGX106005, 2017CXGC0701, and 2017CXGC0706).

## References

1. Albanese, J.S.: Illegal gambling businesses & organized crime: an analysis of federal convictions. *Trends Organ. Crime* **21**(3), 262–277 (2018)
2. Banks, J.: Edging your bets: advantage play, gambling, crime and victimisation. *Crime Media Cult.* **9**(2), 171–187 (2013)
3. Banks, J.: Internet gambling, crime and the regulation of virtual environments. In: Banks, J. (ed.) *Gambling, Crime and Society*, pp. 183–223. Palgrave Macmillan UK, London (2017). [https://doi.org/10.1057/978-1-137-57994-2\\_6](https://doi.org/10.1057/978-1-137-57994-2_6)
4. Blaszczynski, A.: Online gambling and crime: causes, controls and controversies. *Int. Gambl. Stud.* **15**(2), 340–341 (2016)
5. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: *European Conference on Computational Learning Theory*, pp. 23–37 (1995)
6. Gainsbury, S., Wood, R., Russell, A., Hing, N., Blaszczynski, A.: A digital revolution: comparison of demographic profiles, attitudes and gambling behavior of internet and non-internet gamblers. *Comput. Hum. Behav.* **28**(4), 1388–1398 (2012)
7. Gainsbury, S.M., Russell, A., Hing, N., Wood, R., Blaszczynski, A.: The impact of internet gambling on gambling problems: a comparison of moderate-risk and problem internet and non-internet gamblers. *Psychol. Addict. Behav.* **27**(4), 1092–1101 (2013)
8. Gainsbury, S.M., Russell, A., Hing, N., Wood, R., Dan, L., Blaszczynski, A.: How the internet is changing gambling: findings from an Australian prevalence survey. *J. Gambl. Stud.* **31**(1), 1–15 (2013)
9. Gainsbury, S.M., Russell, A., Wood, R., Hing, N., Blaszczynski, A.: How risky is internet gambling? A comparison of subgroups of internet gamblers based on problem gambling status. *New Media Soc.* **17**(6), 861–879 (2015)
10. McMullan, J.L., Rege, A.: Online crime and internet gambling. *J. Gambl. Issues* **24**(5), 115–116 (2010)
11. Tong, S., Zhang, H., Shen, B., Zhong, H., Wang, Y., Jin, B.: Detecting gambling sites from post behaviors. In: *2016 IEEE 11th Conference on Industrial Electronics and Applications*, pp. 2495–2500 (2016)

12. Wardle, H., Moody, A., Griffiths, M., Orford, J., Volberg, R.: Defining the online gambler and patterns of behaviour integration: evidence from the british gambling prevalence survey 2010. *Int. Gambl. Stud.* **11**(3), 339–356 (2011)
13. Wood, R.T., Williams, R.J.: A comparative profile of the internet gambler: demographic characteristics, game-play patterns, and problem gambling status. *New Media Soc.* **13**(13), 1123–1141 (2011)
14. Wood, R.T., Williams, R.J., Lawton, P.K.: Why do internet gamblers prefer online versus land-based venues? Some preliminary findings and implications. *J. Gambl. Issues* **20**, 235–252 (2007)