Multi-source Data Multi-task Learning for Profiling Players in Online Games

Shiwei Zhao, Runze Wu, Jianrong Tao(✉), Manhu Qu, Hao Li, Changjie Fan
*Fuxi AI Lab, NetEase Games, Hangzhou, China
{zhaoshiwei, wurunze1, hztaojianrong, qumanhu, lihao01, fanchangjie}@corp.netease.com

***Abstract*—Profiling game players, especially potential churn and payment prediction, is of paramount importance for online games to improve the product design and the revenue. However, current solutions view either churn or payment prediction as an independent task and most of the previous attempts only depend on the single data source, i.e., the tabular portrait data. Based on the data of two real-world online games, we conduct extensive data analysis. On the one hand, there exists a significant correlation between the player churn and payment. On the other hand, heterogeneous multi-source data, including player portrait, behavior sequence, and social network, can complement each other for a better understanding of each player. To this end, we propose a novel Multi-source Data Multi-task Learning approach, named MSDMT, to capture the multi-source implicit information and predict the churn and payment of each player simultaneously in a multi-task learning fashion. Comprehensive experiments on the two game datasets validate the effectiveness and rationality of our proposed method, which yields significant improvements against other baseline approaches.**

***Index Terms*—player profiling, multi-source data, multi-task learning, online games**

I. INTRODUCTION

The games industry is booming with continued stable revenue more than 151 billion U.S. dollars[2], which has emerged as a promising comprehensive market, more than just an entertainment business. With the widespread popularity of game technology, online games gain a very wide audience and are loved by players of all ages. As a critical component in online games, player profiling has attracted increasing attention from both academic and industrial fields [1]–[3]. A variety of platforms and services centered on player profiling have been deployed in most online games.

Player profiling aims to understand who the players are and what they will do, especially whether they will quit, i.e., the *churn*, and how much they will pay, i.e., the *payment*. The player churn rate largely determines the life of an online game while the potential payment can measure the profit of each game. Combining the churn and payment prediction, game analysts can project the lifetime value (LTV) of each player and the total revenue of the whole game. A wide range of approaches have been developed for either the churn or payment prediction. Generally, the churn prediction is modeled as a classification task [4]–[7] and the payment prediction can be viewed as a regression problem [8], [9]. Statistical predictive methods, e.g., logistic regression [4], tree-based model [5], often depend on the delicate feature engineering while neural network models, e.g., Multi-Layer Perceptron (MLP) [5], Long Short-Term Memory (LSTM) [7], try to explore the implicit correlation given the large-scale data.

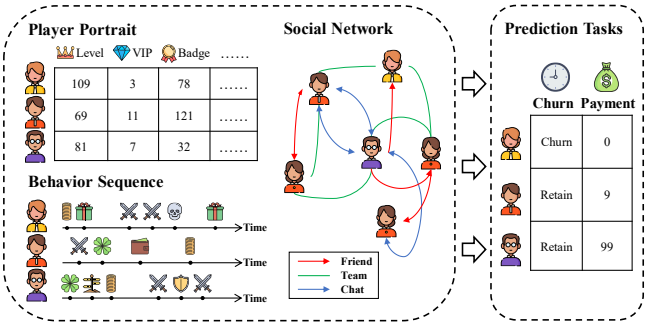


Fig. 1. Illustration of churn and payment prediction based on three heterogeneous data sources in online games. The player portrait describes some players' attributes. The behavior sequence expresses the event type and time for each in-game action of players, where different icon represents different event. And the social network shows the diverse interactions between players.

Despite impressive progress, the current solutions can not fully exploit the correlation between the player churn and payment as well as the complementary between the heterogeneous multi-source data. As shown in Fig. 1, various kinds of gameplay and settings in online games provide multiple sources of heterogeneous data to describe each player: 1) player portrait generated from player behavior records statistics; 2) behavior sequence [10] for every in-game action of each player; 3) social network formed by relationships between players. The player portrait and behavior sequence represent the static and dynamic individual preferences of the player, respectively. Besides, player behavior is often influenced by others in relation because groups interact with each other and share similar event streams, especially at the regular time, places, or relationships. However, most of the previous approaches mainly focus on the tabular player portrait with little consideration of the complementary information brought by the behavior sequence and social network. For example, potential patterns in behavior sequences and group effects of social networks can profile players from different perspectives and complement each other.

*NetEase Fuxi AI Lab: named after Fu Xi, the legendary creator in China, and established to enlighten games with artificial intelligence. (https://fuxi.163.com/en/)

[2]https://newzoo.com/insights/trend-reports/newzoo-global-games-market-report-2019-light-version/

978-1-7281-4533-4/20/$31.00 ©2020 IEEE

Moreover, most of the existing methods in the literature try to predict either the player churn or the player payment as two independent tasks. Intuitively, the LTV of each player is basically determined by how long the player will play and how much the player will pay so that it is very reasonable to consider the churn and payment simultaneously. Empirical analysis (see Sec. IV) shows that there exists a significant correlation between the two tasks, e.g., a player is very unlikely to pay without the desire to play the game. Current single-task solutions are unable to capture the correlation between the player churn and payment for better profiling players. In all, we lack an effective approach that can handle both the payer churn and payment prediction tasks by fully utilizing the multi-source heterogeneous data.

To address these issues, we propose a novel Multi-source Data Multi-task Learning approach, named MSDMT[3], for profiling players with both player churn and payment prediction in online games. To be specific, we build three modules based on player portrait data, behavior sequence data, and social network data to capture rich implicit information from three different perspectives. We employ LSTM to model the dynamic in the daily aggregated player portraits. Considering the impact of potential temporal information, we leverage a hierarchical Convolutional Neural Network-LSTM (CNN-LSTM) to exploit the short- and long-term players' behavioral preferences. Combining the representations of player portrait and behavior sequence, we can build the individual preferences of each player and obtain the node feature in the social network graph. Adopting Graph Convolutional Network (GCN) to mine the group preferences among players, we also optimize the player churn and payment prediction tasks by using a multi-task learning framework. The main contributions are summarized as follows:

- Via extensive empirical observations and analysis, we discover the significant correlation between the player churn and payment as well as the difference and complementary in various heterogeneous data sources.
- Inspired by the key findings, we propose a three-module framework to handle the multi-source data and make final predictions in a multi-task learning fashion.
- We conduct comprehensive experiments on two real-world datasets to verify the effectiveness and the reasonability of our proposed MSDMT.

II. RELATED WORK

A. Profiling Players in Games

Analysis and modeling of player portraits over the lifetime of game players is a widely concerned issue. Players have different needs at different lifetime stages. Therefore, collecting and managing data on the player lifetime will help games to understand players better and create better in-game experiences [11]. To this end, game companies invest significant resources in profiling players, especially for churn and payment. In terms of churn prediction, engagement [12], [13] and retain [14] prediction can also be grouped into the same problem in the game industry, which aims to discover players with inactive or churn intention in the early stage, dig up the reasons and retain them by interventions. Many methods have been proposed by various researchers to address churn prediction in games. Most of these works focus more on extracting salient features from game log data and model it as a binary classification problem by traditional classifiers and neural networks, such as supervised learning [5] and semi-supervised learning [15]. And some researchers have tried to solve it by survival analysis in [6], [12]. Besides, external in-game information of players has also been fused for better churn prediction. Kristensen et al. [7] combine sequential and aggregate data using different neural network architectures for churn prediction in casual freemium games.

Another important aspect of player value is the in-game purchasing power. The ability to effectively predict how much players will pay can help game companies better understand the changing needs of players, thereby more specifically increasing the lifetime value of each player. The literatures related to payment prediction in games can be learned in terms of player lifetime value [8], [9] and player purchasing [16], [17]. Another example of payment prediction is the transition of players between non-paying and premium [18], [19]. At present, techniques and applications based on complex multi-source data fusion are lacking in games, and most of the studies related to game players profiling are still based on single-source data, without taking full advantage of the rich information between various data. In addition, most research only concentrates on either churn or payment prediction, ignoring the correlation between the two tasks.

B. Multi-source Data Fusion

Recently, many works show the strength of multi-source data fusion methods on solving prediction problems in various fields, such as transportation [20], [21], environment [22]–[25], and operations research [26]–[28], to name a few. To complement each other and improve the representational ability of data, multi-source data fusion extracts features from heterogeneous data and fuses them at the data-wise or model-wise, instead of just modeling from single data source. In the field of transportation, considering the effects of spatial-temporal data, multi-source data fusion is mainly based on the raw sensor data of vehicles, roads and pedestrians obtained from infrastructure sensors or probes, and external data (e.g., meteorological, social media, GPS, and event data), and used to solve congestion [20] and flow [21] prediction. For environment, spatial (e.g., road network, POIs, and pollutant distribution) and temporal (e.g., meteorology, traffic, and human mobility) data is fused by the deep distributed fusion network [22], semi-supervised inference model [23] or co-training framework that consists of two separated classifiers [24] to predict air quality. Besides, social media data is also taken into account in [25]. As for operations research, multi-source data fusion is applied in bike sharing systems [26], dispatching demands prediction [27] and price prediction [28] by fusing the implicit

[3]Code available here: https://github.com/fuxiAIlab/MSDMT

TABLE I
DETAILED DESCRIPTION OF THE DATASETS.

Dataset	Player Portrait		Behavior Sequence		Social Network				Samples Statistics		
	#Feature	Level	#Behavior	Avg. Seg. Len.	Graph	Graph Type	#Node	#Edge	#Total	#Churn	#Payment
ACT	79	day	326	272.79	chat	directed weighted	32843	340414	32843	6859	7878
					friend	directed unweighted	32843	281969			
					team	undirected weighted	32843	2390068			
CCG	39	day	229	179.42	guild	undirected unweighted	28380	1538106	28380	7897	5484

features extracted from various data sources. Multi-source data fusion has shown superiority in various scenarios while rare efforts like this are cast into profiling players. In this paper, we also adopt the multi-source data to model each player instead of the single-source data.

III. DATASET

For experimental work reported in this paper, we use two real-world datasets with different game types from NetEase Games[4], including action game (ACT) and collectible card game (CCG). We briefly show the basic statistics of the two datasets in Table I.

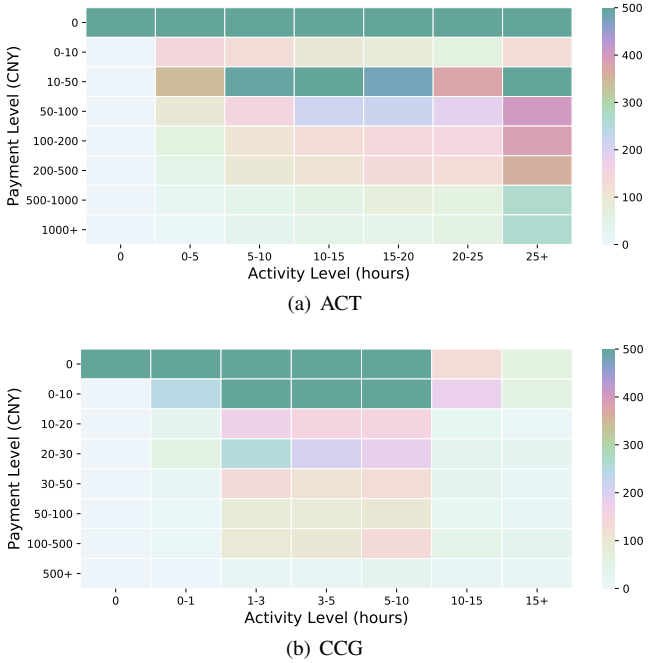
- **ACT.** The ACT dataset describes the attributes, actions, and interactions of players from one mobile action game, named ***Butterfly Sword*** [5],where players can freely experience its challenging Player versus Environment (PvE) and Player versus Player (PvP) gameplay. We conduct player portrait features from several aspects: basic attributes, social relationships, behavioral habits, and consumption preferences. And we built player relationships through chat, friend and team behavior between players.
- **CCG.** The CCG dataset collects the activities and behaviors of players from one mobile collectible card game, named ***Love is Justice*** [6]. During the game, players can interact with Non-Player-Controlled Characters (NPCs) through animated tasks to manage love relationships. Because of the simple single-player gameplay, we only consider the basic attributes, behavioral habits, and consumption preferences as the player portrait features. And we built player relationships through guild in the game.

Both ACT and CCG datasets contain three data sources, i.e., player portrait, behavior sequence, and social network, in which the player portrait is aggregated daily and the behavior sequence is segmented by day from 8 a.m. to 8 a.m. the next day (considering the daily routine of most normal players). For each game, we select players who have logged in from November 29, 2019 to December 1, 2019 as target players. The label window which determines the churn label and payment label in the dataset is from December 2, 2019 to December 8, 2019. We define the player churn if a player is inactive for 7 days and the payment label of a player as the total amount paid by the player for 7 days. For player portrait and behavior sequence, we collect the data from November 25, 2019 to December 1, 2019. As for the social network, we mainly focus on the social relationships between players in the last 180 days, from June 1, 2019 to December 1, 2019.

[4]https://game.163.com/

[5]https://leihuo.163.com/en/games.html?g=game-7#sy

[6]https://leihuo.163.com/en/games.html?g=game-2#nsh



(a) ACT

(b) CCG

Fig. 2. Heatmaps of activity level and payment level for players. Colors represent the number of players for the corresponding level.

IV. ANALYSIS

We point out the relationship between the two tasks of churn and payment first. According to the definitions, we use the total online time and the total recharge amount in the label window to represent the activity level (i.e., churn level) and payment level of the player, respectively. We visualize the distribution of players per activity level and payment level by heatmap in Fig. 2. The results demonstrate that churn and payment do interact and complement each other. First, churn and payment are mutually exclusive because the intersection of the two sets is empty. Then players who do not pay distribute randomly at different levels of activity with no specific pattern (a large number of players gather here), while paying players are more concentrated in high activity levels, which shows that players who are willing to pay will be more active in the future than those who are not. Thus, we can eventually divide the target players into three groups, i.e., churn players, retain & non-paying players, and retain & paying players. Then we conduct data analysis from the following aspects:

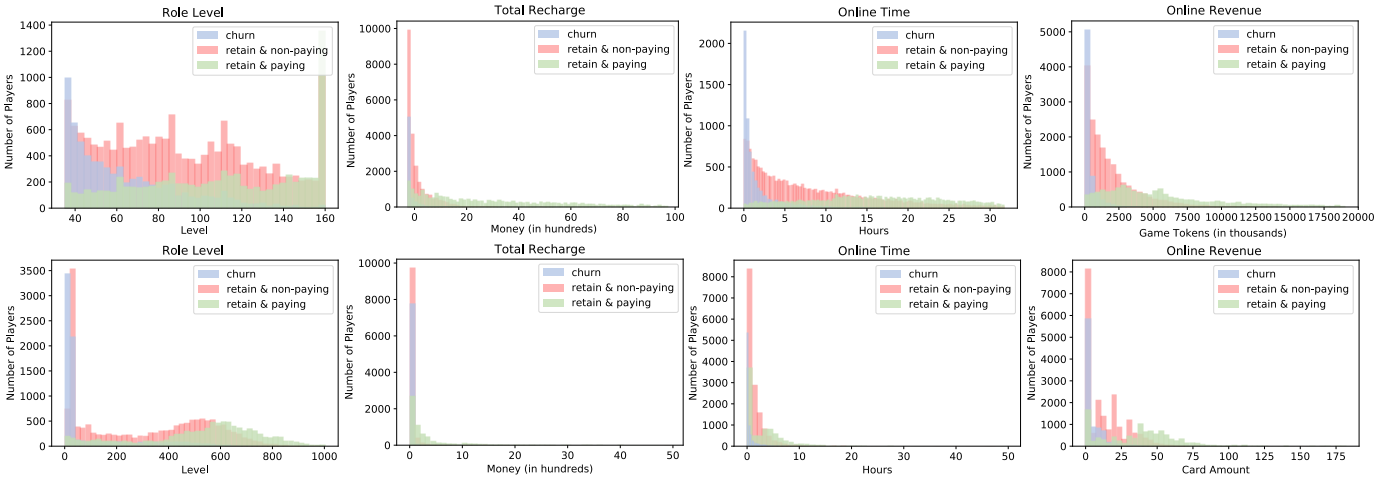


Fig. 3. Histograms of (from left to right) role level, total recharge, online time and online revenue for players, where players with different groups shown in different colors. The top part shows the data from the ACT, and the bottom part shows the data from the CCG.

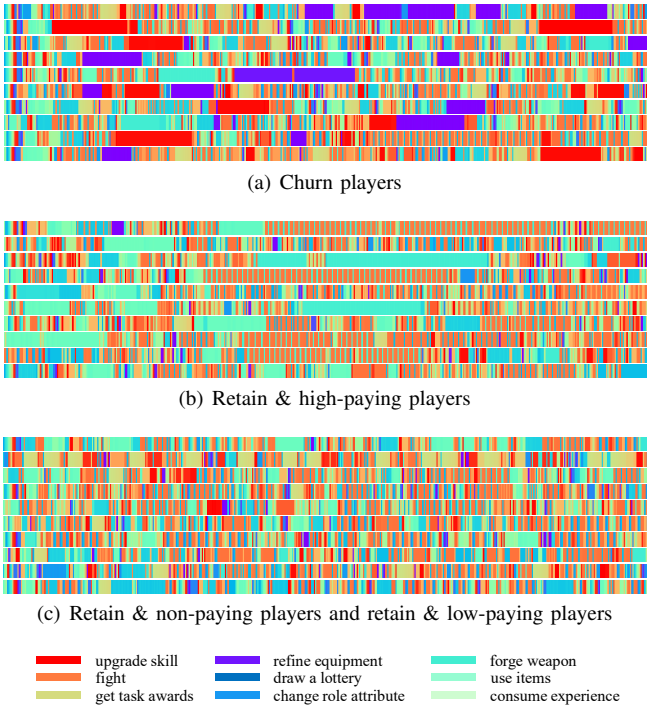
A. Difference in Player Portrait

Fig. 3 shows the distribution histograms of players corresponding to the role level, total recharge, online time, and online rewards (i.e., game tokens obtained in ACT and amount of cards rewarded in CCG), which represent the familiarity, payment potential, game investment, and game revenue of players, respectively. From the results, we can draw the following conclusions: 1) low-level players tend to churn (a lot of players churn even when they just start games), who may join the game just for some activity or novelty. For retained players, high-level players are more willing to pay for a better in-game experience. 2) The amount of a player's historical total recharge represents the player's potential ability to pay and long-term attitude towards the game. Players who have recharged a large amount rarely leave the game, and moreover, they are most likely to continue to recharge. 3) The players' investment and revenue ratio in the game also directly affect the player churn and payment. Players who have spent a lot of time in the game hope to get a fair return, otherwise they will gradually lose interest in the game and pay less. And high revenue drives players to retain and encourages them to pay.

B. Difference in Behavioral Preferences

As shown in Fig. 4, we visualize the behavior sequences of some typical players in ACT[7] to compare the behavioral differences between groups, which we can not learn from the player portrait intuitively. The results reveal some interesting findings as follows: 1) Players always do a lot of continuous operations to consume their previously stored experience and game tokens before churn, such as purple corresponding to refining the equipment, cyan corresponding to forging the weapon, and red corresponding to upgrading the skill in Fig. 4(a). In addition, we count the top five behaviors with the highest frequency by filtering out regular behavior, i.e., settlement rewards, lottery,

[7]We get a similar result on CCG, but here we do not give it due to the limited space.



(a) Churn players

(b) Retain & high-paying players

(c) Retain & non-paying players and retain & low-paying players

Fig. 4. The visualization shows how behavior sequences differ from different groups, where different behavior shown in different colors. To further illustrate the behavioral preferences of the retain & paying player, we visualize the high-paying players among the retained players separately.

reserved experience conversion, daily or main task awards, and gift redemption, which also shows players tend to empty their accounts before they decide to leave the game. 2) The top five most frequent behaviors of retain & high-paying players are lottery, fight, guild activity, chat, and purchasing. On the one hand, high-paying players always stay constantly active in the game and remain keen on multiplayer competitive gameplay, which is corresponding to the orange stripes parts in Fig. 4(b). To improve their competitiveness, they also spend a lot of time forging competitive weapons (corresponding to the cyan

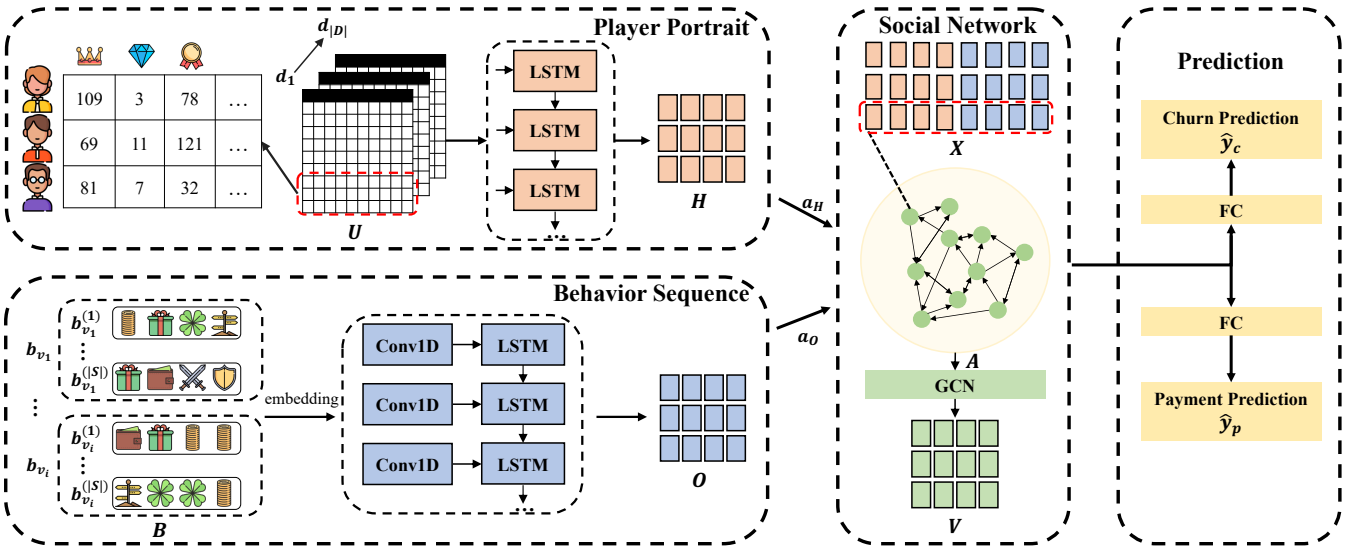
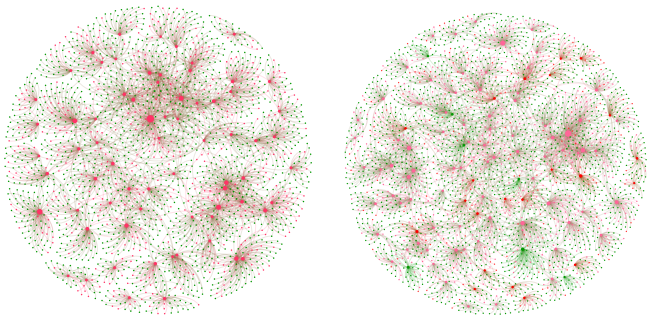


Fig. 5. The framework of our proposed MSDMT.

parts). On the other hand, demand for social and purchasing drives players to continue recharging. 3) Behavior sequences of retained players who do not pay or pay in small amounts are diverse and irregular with randomness.



(a) Churn (b) Payment

Fig. 6. The visualization of social network based on friend relationship in ACT. (a) Red represents churn players and green represents retain players. (b) Green represents non-paying players, while gradation of red indicate the different payment level players.

C. Difference in Group Preferences

Group preferences lead us to profile players from another perspective. Players with similar preferences tend to have relationships and interact with each other, e.g., the top players always like to take on more difficult tasks together, and players are more willing to team up with others they know or buy goods recommended by others they know. We illustrate group differences with the example of the friend relationship from ACT in Fig. 6, which shows that churn players will naturally cluster together to form many small local groups. Players' in-game participation is affected by the players around them. If most of the friends around a player are churn, the player is also likely to churn soon. While if most of the friends around a player are active, there is a high probability that the player will remain active. Similarly, players with different payment levels also tend to cluster and influence each other.

V. METHODS

In this section, we elaborate on the design of our proposed MSDMT as shown in Fig. 5. First, we describe a conceptual definition of the research problem. And then we illustrate the details of three modules built to process different data sources. Finally, we conduct a multi-task learning framework to train and predict both the churn and payment tasks.

A. Problem Definition

Given the three different data sources: 1) player portrait $\mathbf{U} = \{u_{v_i}^{(d)} | v_i \in \mathcal{V}, d \in D\}$; 2) behavior sequence $\mathbf{B} = \{b_{v_i} | v_i \in \mathcal{V}\}$; 3) social network $\mathbf{G} = (\mathcal{V}, \mathcal{E})$ with nodes $v_i \in \mathcal{V}$ and edges $(v_i, v_j) \in \mathcal{E}$, where $\mathcal{V}$ is the set of the players and D denotes the set of day observed the data, our goal is to predict whether the player will churn (i.e., $\hat{y}_c$) and how much the player will pay (i.e., $\hat{y}_p$). In practice, we define the churn prediction task as a binary classification problem and the payment prediction task as a regression problem.

B. Player Portrait Module

In most cases, the player portrait is mainly represented as static features of the players' individual state information without external processing. As shown in the left top part of Fig. 5, considering the data is aggregated by day from the raw game log in our games, we use the player portrait $u_{v_i} \in \mathbf{U}$ as a fixed length sequence, i.e., $u_{v_i} = \{u_{v_i}^{(1)}, u_{v_i}^{(2)}, \cdots, u_{v_i}^{(|D|)}\}$ for each $v_i \in \mathcal{V}$, and adopt LSTM based methods to capture the implicit temporal information of historical player portrait sequences as follows:

$$\mathbf{H} = \text{LSTM}(\mathbf{U}). \tag{1}$$

Here $\mathbf{H}$ is the embedding vector of player portrait module.

C. Behavior Sequence Module

Behavior sequence records the in-game actions of each player, from which we want to model the potential individual behavioral patterns as shown in the left bottom part of Fig. 5. Taking into account the continuity of behavior during in-game time, we preprocess the raw full sequence $b_{v_i} \in \mathbf{B}$ to several segments, i.e., $b_{v_i} = \{b_{v_i}^{(1)}, b_{v_i}^{(2)}, \cdots, b_{v_i}^{(|S|)}\}$ for each $v_i \in \mathcal{V}$, where $|S|$ means a fixed number of segments (we often split segments by days in practice). Specifically, $b_{v_i}^{(s)} = \{(t_1, e_1), (t_2, e_2), \cdots\}$ denotes an event stream in a segment sequence, where a specific type of event e happened at time t. We conduct a hierarchical neural network to embed the behavior sequence. First, we encode the each segment sequence by one-dimensional convolutional neural network (Conv1D). Then we stack the output (i.e., join a sequence of arrays along a new axis) to form a new sequence and feed it into LSTM as follows:

$$\begin{aligned} \hat{\mathbf{B}} &= \text{Stack}(\text{Conv1D}(\mathbf{B}^{(\mathbf{s})})), \text{ for } s \in S, \\ \mathbf{0} &= \text{LSTM}(\hat{\mathbf{B}}). \end{aligned} \tag{2}$$

Here $\mathbf{O}$ is the embedding vector of behavior sequence module, S denotes the set of segments and $\hat{\mathbf{B}}$ represents the stack output of hidden representation of each segment.

D. Social Network Module

We consider player portrait and behavior sequence together to capture individual preferences of the players more comprehensively, and we also take into account the group preferences between players. To this end, we leverage GCN [29] to fully combine the complementary information between the three heterogeneous data sources as shown in the middle part of Fig. 5. Given embedding vectors of player portrait $\mathbf{H}$ and behavior sequence $\mathbf{O}$, and graph adjacency matrix $\mathbf{A}$, we perform an attention based method to obtain weights α_H and α_O to concatenate $\mathbf{H}$ and $\mathbf{O}$, which will be used as the node feature with the $\mathbf{A}$ together to be fed into GCNs:

$$\begin{aligned} \mathbf{A} &= \begin{cases} 0 & (v_i, v_j) \notin \mathcal{E} \\ 1 & (v_i, v_j) \in \mathcal{E} \end{cases}, \\ \mathbf{X} &= \text{Concat}(\alpha_H \mathbf{H}, \alpha_O \mathbf{O}), \\ \mathbf{V} &= \text{GCN}(\mathbf{X}, \mathbf{A}). \end{aligned} \tag{3}$$

Here $\mathbf{V}$ is the final fused embedding vector and $\mathbf{X}$ represents the fused node feature vector.

E. Multi-task Learning

With the constructed above modules, we build the training and prediction module based on multi-task learning. We consider two different losses for two prediction tasks, i.e., Cross-entropy Loss for churn task and Mean-Squared Error Loss for payment task. To be specific, y_c, $\hat{y}_c$ correspond to the label and prediction for churn prediction task, and y_p, $\hat{y}_p$ correspond to the label and prediction for payment prediction task, we compute a loss function that jointly evaluates the performance of all tasks, which can be expressed as follows:

$$\mathcal{L} = \sum_{v_i \in V} (-\alpha y_c \log(\hat{y}_c) - \beta (y_p - \hat{y}_p)^2). \tag{4}$$

Here the loss $\mathcal{L}$ is over all the training data and α, β are hyperparameters that balance the weighting. In practice, it is non-trivial to keep two losses at the same order of magnitude during training, thus none of them would dominate in gradient computation. To this end, the log transformation is used to address skewed distributions in payment data. And we set the α to 0.7 and β to 0.3.

VI. EXPERIMENTS

In this section, we introduce the experiments on two real-world datasets. First, we show the experiment results of MSDMT compared against other baseline methods. Besides, we conduct ablation study to verify the effectiveness of MSDMT.

A. Baselines

For different data sources, we choose several competitive methods as baselines to show the performance of single data source in both churn and payment prediction tasks. Considering player portrait is aggregated by day in our datasets, just like sequence data, we choose the following baseline methods:

- **LSTM:** LSTM is a classical architecture specifically designed for sequence prediction problems with spatial or temporal inputs. Here we employe LSTM to model potential dependencies over the player portrait sequence.
- **CNN:** CNN is also widely used to processing sequence data, and we use Conv1D to extract features of the player portrait sequence in our experiments.

For behavior sequence, we compare the performance of LSTM and CNN to extract features from segmented sequences and combine them for sequence prediction as follows:

- **LSTM:** We leverage a hierarchical LSTM to extract the features of each segmented sequences respectively and combine them to make sequence prediction.
- **CNN-LSTM:** The architecture of CNN-LSTM involves using CNN layers (Conv1D for sequence inputs) for feature extraction on input data combined with LSTM to perform sequence prediction on the feature vectors.

In addition, we compare the GCN with different relationships without external node feature (i.e., only use the identity matrix as the node feature) to demonstrate the effectiveness of the graph structure. We select the relationship with the best performance as the input of social network module in our proposed method for each game. Lastly, MSDMT-single and MSDMT-multi represent MSDMT without multi-task learning and with multi-task learning, respectively. We comprehensively compare the performance of baseline models on each data sources to verify the effectiveness of multi-source data fusion and multi-task learning in our proposed method.

TABLE II
PERFORMANCE COMPARISON AMONG DIFFERENT METHODS IN VARIOUS DATA SOURCES.

Data	Method	ACT					CCG				
		Churn Task			Payment Task		Churn Task			Payment Task	
		ACC	AUC	F1-Score	RMSE	MAE	ACC	AUC	F1-Score	RMSE	MAE
Player Portrait	LSTM	0.8606	0.9065	0.6526	99.2221	26.5535	0.8806	0.9466	0.7897	32.0225	6.1233
	CNN	0.8531	0.8976	0.6136	102.5784	27.6703	0.8855	0.9439	0.7939	33.4217	6.3421
Behavior Sequence	LSTM	0.8598	0.9031	0.6482	193.8058	33.4812	0.8698	0.9314	0.7869	163.2042	9.3127
	CNN-LSTM	0.8645	0.9078	0.6731	110.2557	30.4499	0.8829	0.9457	0.7923	139.2158	8.5813
Social Network	GCN-chat	0.6544	0.6788	0.3985	596.2772	49.4030	-	-	-	-	-
	GCN-friend	0.6894	0.7021	0.4344	518.7393	45.6390	-	-	-	-	-
	GCN-team	0.6737	0.6909	0.3729	521.5121	47.3167	-	-	-	-	-
	GCN-guild	-	-	-	-	-	0.7565	0.8034	0.4647	323.2026	13.4812
Multi-source data	MSDMT-single	0.8705	0.9177	0.6747	96.3817	**26.5148**	0.8944	0.9605	0.8193	29.6693	6.0151
	MSDMT-multi	**0.8742**	**0.9245**	**0.6775**	**94.3422**	26.5679	**0.8974**	**0.9642**	**0.8204**	**29.1409**	**5.8625**

TABLE III
ABLATION EVALUATION OF DIFFERENT MODULES PROPOSED IN OUR METHOD.

Module	ACT					CCG				
	Churn Task			Payment Task		Churn Task			Payment Task	
	ACC	AUC	F1-Score	RMSE	MAE	ACC	AUC	F1-Score	RMSE	MAE
Player Portrait	0.8623	0.9092	0.6462	96.1658	27.0679	0.8889	0.9459	0.8051	31.8503	6.3747
Behavior Sequence	0.8601	0.9055	0.6639	145.5441	31.4512	0.8761	0.9462	0.8006	124.2308	8.1399
Social Network	0.6812	0.6992	0.4310	510.1241	44.9654	0.7681	0.8129	0.4771	312.7412	13.0112
Player Portrait + Behavior Sequence	0.8689	0.9142	0.6642	96.4213	**26.0144**	0.8927	0.9565	**0.8263**	31.2109	6.1625
Player Portrait + Social Network	0.8611	0.9127	0.6572	97.1214	26.8745	0.8901	0.9517	0.7973	32.7901	6.7523
Behavior Sequence + Social Network	0.8645	0.9085	0.6631	143.0174	32.1247	0.8847	0.9481	0.8133	114.1254	7.8415
MSDMT	**0.8742**	**0.9245**	**0.6775**	**94.3422**	26.5679	**0.8974**	**0.9642**	0.8204	**29.1409**	**5.8625**

B. Experimental Settings

We use three widely binary classification evaluation metrics, i.e., Accuracy (ACC), Area Under Curve (AUC), and F1-Score as evaluation metrics on the churn task. And we use Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) metrics to evaluate the different methods on the payment task. For each game, we split 80% of the dataset into a training set and the remaining 20% into a test set. All experiments are repeated 5 times and the average results are reported. For experimental results, the standard error is below 0.001 (AUC), 0.5 (RMSE) and not shown due to the limited space.

In MSDMT, the player portrait module is a classical LSTM layer with 64 hidden units. The CNN-LSTM structure of behavior sequence module has a CNN layer with 64 filters and a LSTM layer with 32 hidden units, where the convolutional kernel is a one-dimension vector and the length of the kernel is set to 32. We employ two GCN layers with 64 hidden units in social network module and the dropout rate is set to 0.5.

C. Performance Comparison

Table II shows the comparison results on both two real-world datasets. From the perspective of data source, social network achieves higher performance than the random benchmark, but performs worse than other single-source data based methods significantly, which demonstrates the effectiveness of node feature and graph structure in graph neural network. Player portrait and behavior sequence benefit from the rich information of their own and perform better both in two tasks. Compared with methods base on single-source data, we propose the effective multi-source data method MSDMT, which can not only consider the difference in heterogeneous data, such as tabular, sequence, and graph, but also exploit the complementary information of multiple data sources. On the other hand, capturing implicit representations of raw data from different data sources is very important for multi-source data fusion. The experimental results show that LSTM and CNN-LSTM perform better in player portrait and behavior sequence, respectively, which verify the rationality of different modules proposed in our method. Furthermore, benefit from multi-source data and multi-task learning, MSDMT achieves 1.84%, 1.86% (AUC) and 4.92%, 9.00% (RMSE) improvement over the best single-source data single-task learning based method on both the churn and payment tasks, respectively. And MSDMT with multi-task learning performs against MSDMT without multi-task learning with an increment of 0.74%, 0.39% (AUC) and a reduction of 2.12%, 1.78% (RMSE).

D. Ablation Study

To further illustrate that the fusion of different data sources can improve the model performance with multi-task learning, we study the effect of the interaction between different modules proposed in our method. Table III shows the performance of MSDMT and its variants. First, we can see that single-source data is not very effective, especially for social network without external node features. In most cases, the experimental results demonstrate the significant effectiveness of interactions between different data sources. Player Portrait + Behavior Sequence achieves a higher AUC (an increment of 0.05, 0.01) and a lower RMSE (a reduction of 1.05, 0.64) over the best single-source data based method on both the churn and

payment tasks, respectively. Furthermore, the graph structure in social network does play an important role both in churn and payment prediction. By fusing the social network, the performance of the variant models is improved. Lastly, the performance is best when all data sources are fused. The more data sources are combined, the more diverse the information is contained, and the more performance model improves. Compared with the method based on single-source data, MSDMT yields 1.68%, 1.90% (AUC) and 1.90%, 8.51% (RMSE) improvement. And MSDMT also performs best against methods fused by two data sources with 1.13%, 0.81% (AUC) and 2.16%, 6.63% (RMSE) improvement.

VII. CONCLUSION

In this paper, a novel approach named Multi-source Data Multi-task Learning (MSDMT) is proposed to profile players in online games. We conduct extensive empirical observations and analysis on the data of two real-world online games. The results establish a significant correlation between the player churn and payment, and also verify the difference and complementary in various heterogeneous data sources. Inspired by the key findings, MSDMT builds three different modules based on heterogeneous data sources such as player portrait, behavior sequence, and social network, to capture and fuse the rich implicit information. Moreover, MSDMT makes the churn and payment prediction simultaneously in a multi-task learning fashion. Comprehensive experiments on the two game datasets show the superiority of MSDMT on both churn and payment prediction. For future work, we plan to further investigate and improve the multi-source data fusion method of MSDMT for better performance. In addition, we also consider expanding our work to more applications in online games.

REFERENCES

[1] A. Canossa, S. Makarovych, J. Togelius, and A. Drachenn, "Like a dna string: Sequence-based player profiling in tom clancy's the division," in *Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2018.

[2] I. Samborskii, A. Farseev, A. Filchenkov, and T.-S. Chua, "A whole new ball game: Harvesting game data for player profiling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 10 025–10 026.

[3] A. F. del Río, P. P. Chen, and A. Periánez, "Profiling players with engagement predictions," in *2019 IEEE Conference on Games (CoG)*. IEEE, 2019, pp. 1–4.

[4] F. Hadiji, R. Sifa, A. Drachen, C. Thurau, K. Kersting, and C. Bauckhage, "Predicting player churn in the wild," in *2014 IEEE Conference on Computational Intelligence and Games*. IEEE, 2014, pp. 1–8.

[5] J. Runge, P. Gao, F. Garcin, and B. Faltings, "Churn prediction for high-value players in casual social games," in *2014 IEEE conference on Computational Intelligence and Games*. IEEE, 2014, pp. 1–8.

[6] E. Lee, Y. Jang, D.-M. Yoon, J. Jeon, S.-i. Yang, S.-K. Lee, D.-W. Kim, P. P. Chen, A. Guitart, P. Bertens *et al.*, "Game data mining competition on churn prediction and survival analysis using commercial game log data," *IEEE Transactions on Games*, vol. 11, no. 3, pp. 215–226, 2018.

[7] J. T. Kristensen and P. Burelli, "Combining sequential and aggregated data for churn prediction in casual freemium games," in *2019 IEEE Conference on Games (CoG)*. IEEE, 2019, pp. 1–8.

[8] A. Drachen, M. Pastor, A. Liu, D. J. Fontaine, Y. Chang, J. Runge, R. Sifa, and D. Klabjan, "To be or not to be... social: Incorporating simple social features in mobile game customer lifetime value predictions," in *Proceedings of the Australasian Computer Science Week Multiconference*, 2018, pp. 1–10.

[9] P. P. Chen, A. Guitart, A. F. del Río, and A. Periánez, "Customer lifetime value in video games using deep learning and parametric models," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 2134–2140.

[10] J. Tao, J. Xu, L. Gong, Y. Li, C. Fan, and Z. Zhao, "Nguard: A game bot detection framework for netease mmorpgs," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 811–820.

[11] R. Flunger, A. Mladenow, and C. Strauss, "Game analytics on free to play," in *International Conference on Big Data Innovations and Applications*. Springer, 2019, pp. 133–141.

[12] M. Viljanen, A. Airola, J. Heikkonen, and T. Pahikkala, "Playtime measurement with survival analysis," *IEEE Transactions on Games*, vol. 10, no. 2, pp. 128–138, 2017.

[13] V. Bonometti, C. Ringer, M. Hall, A. R. Wade, and A. Drachen, "Modelling early user-game interactions for joint estimation of survival time and churn probability," in *2019 IEEE Conference on Games (CoG)*. IEEE, 2019, pp. 1–8.

[14] A. Drachen, E. T. Lundquist, Y. Kung, P. Rao, R. Sifa, J. Runge, and D. Klabjan, "Rapid prediction of player retention in free-to-play mobile games," in *Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2016.

[15] X. Liu, M. Xie, X. Wen, R. Chen, Y. Ge, N. Duffield, and N. Wang, "A semi-supervised and inductive embedding model for churn prediction of large-scale mobile games," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 277–286.

[16] H. Xie, S. Devlin, D. Kudenko, and P. Cowling, "Predicting player disengagement and first purchase with event-frequency based data representation," in *2015 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 2015, pp. 230–237.

[17] S. Harada, K. Taniguchi, M. Yamada, and H. Kashima, "In-app purchase prediction using bayesian personalized dwell day ranking," 2019.

[18] A. Guitart, A. F. del Río, and Á. Periáñez, "Understanding player engagement and in-game purchasing behavior with ensemble learning," *arXiv preprint arXiv:1907.03947*, 2019.

[19] A. Guitart, S. H. Tan, A. F. d. Río, P. P. Chen, and Á. Periáñez, "From non-paying to premium: predicting user conversion in video games with ensemble learning," in *Proceedings of the 14th International Conference on the Foundations of Digital Games*, 2019, pp. 1–9.

[20] L. Lin, J. Li, F. Chen, J. Ye, and J. Huai, "Road traffic speed prediction: a probabilistic model fusing multi-source data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 7, pp. 1310–1323, 2017.

[21] D. Li, W. Wang, M. Liang, and Y. Liu, "Systematic study on the forecasting of transit passenger flow based on machine learning with multi-source data," *Journal of Computing in Civil Engineering*, 2017.

[22] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng, "Deep distributed fusion network for air quality prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 965–973.

[23] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 437–446.

[24] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 1436–1444.

[25] X. Ni, H. Huang, and W. Du, "Relevance analysis and short-term prediction of pm2. 5 concentrations in beijing based on multi-source data," *Atmospheric environment*, vol. 150, pp. 146–161, 2017.

[26] J. Liu, L. Sun, W. Chen, and H. Xiong, "Rebalancing bike sharing systems: A multi-source data smart optimization," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1005–1014.

[27] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li, "Deep multi-view spatial-temporal network for taxi demand prediction," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] C. Ge, Y. Wang, X. Xie, H. Liu, and Z. Zhou, "An integrated model for urban subregion house price forecasting: A multi-source data perspective," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 1054–1059.

[29] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.