Player Analytics: Papers from the AIIDE Workshop
AAAI Technical Report WS-16-23

Predicting Disengagement in Free-To-Play Games with Highly Biased Data

Hanting Xie and **Sam Devlin** and **Daniel Kudenko**
Department of Computer Science
University of York
Deramore Lane, York, YO10 5GH, UK
{hx597|sam.devlin|daniel.kudenko}@york.ac.uk

Abstract

A vital application of game data mining is to predict player behaviour trends such as disengagement, purchase, etc.. Several works have been done by quantitative methods in the last decade. Generally, predicting player behaviour trends is a classification problem where class labels of instances are decided by predefined definitions. However, as the majority of current definitions distribute players into classes only by satisfying specific conditions, a highly biased class distribution may be led to if few (or most) players can satisfy these conditions. In this work, a new definition named *trend over varying dates* that can create balanced class distributions will be introduced and, as an example, disengagement prediction will be used to show how the definition works. Experiments on three commercial mobile games will show how this definition can be applied to games of various genres. Finally, the performance of this definition towards predicting disengagement will be compared with another disengagement concept called 'churn'. Both game-specific and event frequency based data representation (introduced in previous work) will be applied to represent the datasets for predictions. Results indicate that the definition of 'trend over varying dates' can improve the predictive performance by balancing the class distributions in most cases.

Introduction

With the rapid development of game industry, game analytics has become never become so popular. Game companies start to apply this technology during development stage (El-Nasr, Drachen, and Canossa 2013) as it can not only offer a good understanding of players but also help to make important decisions. Predictive modelling is an element of game analytics which is able to provide statistical models of players' behaviours and generate predictions that can help to avoid unnecessary risks (Yannakakis et al. 2013).

Player engagement is a basic behaviour metric that companies would like to address. Several research publications have been done on predicting players' churn/disengagement by modelling their in-game behaviours (Mahlmann et al. 2010; Runge et al. 2014; Weber et al. 2011; Xie et al. 2014; 2015). Although they have provided effective approaches to predict churn/disengagement, many assumed the distribution of disengagement/non-disengagement classes to be even. Generally, disengagement prediction is a classification problem where players need to be split into two classes, i.e., disengagement and non-disengagement. To get disengaging players, most definitions tried to filter disengaging players out by some specific conditions (e.g., recent login time) and consider the rest as non-disengaging players. Although players are labelled in an easy implemented and natural way, the resultant class distribution can be highly biased to one side (either disengagement or non-disengagement) depending on different games and how strict the filtering condition is. An example was observed when we applied the churn definition (introduced by Runge et al.) in the game Race Team Manager for predicting disengagement. A ratio around 4/1 was found between both target classes after the labelling process. According to the experiments conducted on three commercial games, this amount of bias is risky for creating reliable classifiers. A naive solution to balance class distribution is to manually remove samples from the majority class (Chawla 2005). However, if the distribution is highly biased, removing excessive training examples may lead to overfitting. This random sampling method was included in the experiments for comparison.

Another classical predictive task in the game data mining is the purchasing behaviours. Similar to disengagement, biased situations can easily occur (Xie et al. 2015). In most mobile games, especially for free to play ones, it is common that only a few percentage of players would have purchased many items. Due to this, if the purchasing counts were used as the condition for partitioning players into classes, the resultant class distribution will be highly biased, too.

In order to solve this type of problem, this work presents a new general labelling method, named *'Trend Over Varying Dates'*, which is able to maintain an approximately balanced distribution of resultant classes without losing any samples for predicting behaviour trends. Instead of strictly categorising users by setting conditions, this new definition looks at seeking for a soft/dynamic splitting date line that can divide the whole data space into two classes. In this work, we took the popular predictive task disengagement prediction as an example to apply this method to. In this context, the player activities on both sides of the soft/dynamic splitting date will be compared and used as the criteria to distribute play-

ers into either disengagement or non-disengagement class. Our results show this labelling method outperform previous ones (Runge et al. 2014; Chawla 2005) across three commercial games with two different data representation approaches (Runge et al. 2014; Xie et al. 2014).

Related Work

Game Analytics and Data Mining

During the game development process, *game analytics* is a tool that can help to uncover important patterns from game metrics that can support decision-making (El-Nasr, Drachen, and Canossa 2013; Xie et al. 2015). As a subset of game analytics, game data mining applies machine learning technologies for extracting latent patterns or statistical models from massively scaled game metrics (Yannakakis 2012).

Supervised learning, as an important part of machine learning, is designed to build predictive models from labelled datasets (Mohri, Rostamizadeh, and Talwalkar 2012; Xie et al. 2015). *Classification* is a subset of Supervised learning where the labels are nominal. The aim of it is to build up a model that can reflect the correlations between some given target labels (binary or multi) and some selected data representations (also referred to as features). Base on which, in prediction, the resultant model is able to find the correct labels for the unseen examples (Alpaydin 2004). In game-related research works, supervised learning was widely used for predicting players' possible behaviours in the future (Weber et al. 2011). All experiments conducted in this research are supervised learning problems. Thus, classical algorithms such as Decision Tree, Logistic Regression and Support Vector Machine (SVM) have been applied.

Decision Trees

A standard decision tree is a tree-like structure that abstracts features into nodes and forms its branches and leaves following divide-and-conquer strategy (Alpaydin 2010; Xie et al. 2015). As explained by Apt and Weiss, a decision tree is one of the most interpretable model which links every feature (node) to their consequences until reaching the terminal leaves.

Logistic Regression

Logistic Regression is a linear regression model for solving classification problems. It aims at optimising the parameters of a linear model which can correctly describe the relationships between the dependent targets (labels) and selected independent variables (features) (Hosmer and Lemeshow 2004).

Support Vector Machine

A SVM model is built in mainly two steps. At first, it takes the usage of bounded training samples from each class as support vectors that represent them and then maps these vectors into a higher dimension by specified kernel functions (e.g., Gaussian kernel) (Campbell and Ying 2011; Xie et al. 2015). Followed by which, the algorithm seeks for an optimised hyperplane that can maximise the distance between canonical hyperplanes formed by support vectors (Campbell and Ying 2011; Xie et al. 2015).

Churn Prediction

There have been previous efforts focused on predicting players' trends of leaving a game. 'Churn', as a commonly used definition, was mentioned in recent work by Runge et al. and Hadiji et al.. In their works, 'churn' was described as the behaviour that a player entirely stopped his/her activity in games.

According to the work by Runge et al., players who are active and high-valued are churning on a specific day (day 0) if he/she starts 14 consecutive days of inactivity from any days between day 0 and 6. This definition contains three conditions, first, a player to be considered has to be a high-valued player. According to them, a player is said to be high-valued if he/she is in the top 10% of players who are sorted by the revenue generated. Additionally, a player has also to be active enough to be considered. This is defined by observing whether a player played the game at least once between day -14 and day -1. Finally, the last condition takes players who start a 14 consecutive days of inactivity from any days between day 0 and 6 as the churn players. This definition is defined by splitting players with a highly restrictive condition. Due to this, there will be a chance that the resultant two classes (churn and non-churn) are highly biased.

Disengagement Prediction

Disengagement is a similar concept for describing the churning trend of players (Xie et al. 2014; 2015). It also relies on specific conditions to split players into binary groups. The procedure is shown below:

1. For each player, their total activities (the sum of all event frequency features) in both month 1 and month 2 will be calculated separately and sorted.
2. For each month the sorted list of total activities is divided into 4 quartiles and the players are then ranked between 4 and 1 according to which quartile they are within.
3. For each player, if his/her rank in month 1 minus his/her rank in month 2 is greater than 2, then he/she would be allocated to the Disengagement Group. Otherwise, he/she would be allocated to the Non-Disengagement Group.

Because the labelling procedure also relies on strict conditions, disengagement also suffers from the same bias risk as the churn definition does.

Trend Over Varying Dates

In order to solve the problem faced by the current definitions, this work presents a new class labelling approach named **trend over varying dates** that tries to create the most balanced class distribution whilst making use of every data sample in the dataset. Taking disengagement prediction as an example, the class labels are defined by two varying parameters: ***prr*** (prior rounds) and ***por*** (post rounds). The 'prior rounds' stands for the quantities of rounds that a player completed before a splitting date $\boldsymbol{T}$ (T may vary for different players) whilst the 'post rounds' represents the number of rounds that he/she finished after that date.

Note that *prr* and *por* are parameters that need to be manually decided whereas T will simply be the date when player

finished *prr* rounds of games. Based on this, a player is considered as disengaging if he finished *prr* rounds before some *T* but is unable to finish *por* rounds afterwards. Equation 1 defined this method formally. From the perspective of definition, different from traditional churn which aims at describing players who is entirely leaving the game, this disengagement over varying dates approach stresses on detecting the disengaging trends. Regarding its meaning in practical applications, after the class distribution are balanced by optimising the pair of *prr* and *por*, (nearly) half of players who played *prr* rounds of games are still interested in playing another *por* rounds whereas the other half are not. Furthermore, the resultant *prr* can be seen as an indicator of the game's health with regard to retention. Because players are evenly split into disengaging and engaging groups after the date T, a higher *prr* shows that the game keeps players engaged for longer, i.e., most players have played many (*prr*) rounds before half of them will show a disengaging trend. On the other hand, a lower *prr* indicates that half of players start to display a disengaging trend after only a few plays. This suggests that a negative first impression of the game is an important factor in discouraging players from continuing to engage with the game. Additionally, *por* indicates how long a company has to prevent players disengagement by attempting an intervention (e.g. offering in-game bonuses). A bigger *por* means that most players are still able to play many rounds of games after *T* and before disengaging, whilst on the contrary, a small *por* means most players will disengage soon after *T*.

$$\text{total} = total\ rounds\ played$$

$$\text{player label} = \begin{cases} disengaging, & \text{if } total - prr < por \\ engaging, & \text{otherwise} \end{cases} \quad (1)$$

In order to work out the best combination of *prr* and *por* that can balance the disengagement and non-disengagement class, this work applied a standard genetic algorithm (5000 generations, 10 candidates and 0.5 mutation rate) for gaining the smallest distance between two classes.

This method is temporarily referred to as 'disengagement over varying dates' for the rest of this work because it is for disengagement prediction. For other biased predictive purposes, one can easily apply the same equation but replacing *prr* and *por* with other corresponding information instead. For example, to predict purchase behaviour trend, these two parameters can be changed to the purchasing behaviour counts before and after the varying splitting date.

Data Sources

Data that has been used in this research are from three different commercial games, including a professional football player simulation game named ***I Am Playr***, a racing game named ***Race Team Manager*** and a music game called ***Lyroke***. Both ***I Am Playr*** and ***Lyroke*** were developed by ***WeR Interactive*** whilst ***Race Team Manager*** was produced by ***Big Bit Ltd***.

Race Team Manager

Race Team Manager is a free to play game developed by Bit Bit Ltd and it is available across all mobile platforms. The game was picked as the 'Editor Choice' after its first launch on the *App Store*. Its gameplay allows players to take the role of manager of a team who could control how the racing cars should drive to perform overtaking, avoid collisions, reduce tires replacing time and adjust driving styles. In this experiment, the dataset used was full gameplay logs of 113872 players between October, 2015 and January, 2016.

I Am Playr

Developed by WeR Interactive, I Am Playr is another commercial published free to play on multiple platforms. This is a game about football simulation. In the game, user will act as a professional football player and experience his life. The game offered several different actions such as playing league matches, finishing daily trainings and attending special events. During a football match, scrolling text was used for describing the status of matches until a shooting chance is given for player to score a goal. In the present experiment, the dataset used was full gameplay logs of 89057 players during January and February, 2014.

Lyroke

The name of Lyroke comes from the word 'lyrics' as it is a game about guessing lyrics. This game is available to be played across multiple platforms. In the game, players can choose to either play in a tournament mode or challenge their friends. As for the gameplay, once a song stops in the middle, players need to select the next word in the lyrics from possible options that pop up. In this work, the dataset used from Lyroke was full gameplay logs of 280338 players during March and April, 2014.

Pre-processing For Simulating Problem

As mentioned in the Introduction Section, this work was firstly inspired by the dataset from Race Team Manager. It is because the ratio between churn and non-churn shows a highly biased distribution when the original churn definition was applied. Due to this characteristic, the dataset is naturally suitable for being labelled with the new disengagement over varying dates definition (using equation 1).

Unlikely, the class distribution of I Am Play and Lyroke are close to balanced when the original churn was used as the filter. Normally, regarding this, it shall be needless to apply disengagement over varying dates for balancing the distributions. However, in order to show the generality of our new labelling method and perform similar experiments, this work manually simulated the highly biased situation by firstly labelling the data with the original churn and then randomly removing several examples until the class distribution are similar to the Race Team Manager dataset. Because of this, since that the original churn definition only focuses on high value and active players, after applying the filter, the number of player have been decreased a lot for both games.

Methodology

Prediction with Game Specific Features

This research firstly tries to predict the disengagement over varying dates with features similar to those mentioned in the work by Runge et al.. The features picked by their works are 'Rounds played', 'Accuracy', 'Invites sent', 'Days in game', 'Last purchase' and 'Days since last purchase' which covered both players' engaging behaviours and purchasing behaviours. They are all summarised from the raw data by pre-processing. Since 'Rounds played' is part of the definition of disengagement over varying date, it was not selected as a valid feature for our experiments. Apart from that, as discussed in previous research (Xie et al. 2015), since some of these features are not available for some types of games, it is not feasible to find all of them in some of our three games. More precisely, 'Accuracy' is not available for I Am Playr whilst both Lyroke and Race Team Manager are lack of the feature 'Days in game'.

Prediction with Event Frequency Based Data Representations

The definition of event frequency based data representations was first mentioned in our previous research(Xie et al. 2014) for predicting disengagement. It uses only counts of occurrences of events (e.g., win/lose a match, shoot at goal) regardless of their meanings to form the input feature space. This data representation is more general than most other state of the art methods because it does not rely on any specific information of the game to be applied. It has achieved competitive results for predicting both disengagement and churn in previous works (Xie et al. 2014; 2015; Runge et al. 2014). In this paper, similar to game specific features, event frequency based data representation was also used as one data representation method for predict disengagement over varying dates.

Classification Algorithms

In this paper we applied three different machine learning algorithms for this classification problem, i.e., Decision Tree (criterion = 'entropy', splitter = 'best', max_features = None, max_depth = None), Logistic Regression (penalty (the norm) = 'L2', C (Inverse of regularization strength) = 1.0) and Support Vector Machine (C (Penalty parameter) = 1.0, kernel = 'linear', gamma = 1/N features, coef0 = 0.0). The experiments utilised implementations of them from python machine learning package named 'sci-kit learn' (version 0.17.1) with their default parameters.

Evaluation

To make sure overfitting is avoided, results shown in this work are mean values from 10-fold cross validation. For measuring the performance of models, area under ROC (Receiver Operating Characteristic) was used as the first indicator. A ROC is a curve formed by the 'true positive rate' and the 'false positive rate' of the classifier(Davis and Goadrich 2006) which is widely applied in machine learning for evaluating predictive performances. A similar measurement called PRC is formed by 'precision' and 'recall' instead. According to Davis and Goadrich, the ROC considers the performance of model for predicting both positive and negative classes whereas PRC only focuses on the performance for predicting positive class, i.e., regardless of true negatives. Due to this, PRC is often affected by biased datasets. Same situation happens when applying another measurement named F-Measure as it only considers the accuracy for predicting positive examples, too (Powers 2011). Different from both of them, Jeni, Cohn, and De La Torre claimed that ROC is not affected by biased dataset. This is the reason that PRC or F-Measure was not used as one of the measurements.

Cohen's Kappa

$$
\begin{aligned}
&total = tp + tn + fp + fn \\
&positive_{actual} = \frac{tp + fn}{total} \\
&positive_{predicted} = \frac{tp + fp}{total} \\
&negative_{actual} = \frac{fp + tn}{total} \\
&negative_{predicted} = \frac{fn + tn}{total} \\
&\boldsymbol{p_e} = positive_{actual} \cdot positive_{predicted} \\
&+ negative_{actual} \cdot negative_{predicted} \\
&\boldsymbol{p_o} = \frac{tp + tn}{total} \quad \boldsymbol{k} = \frac{p_o - p_e}{1 - p_e}
\end{aligned}
\tag{2}
$$

Similar to ROC, another widely used statistical measurement for biased situation is called Cohen's Kappa. Originally proposed for a different purpose, Cohen's Kappa was firstly introduced by Cohen for calculating the inter-raters agreements (Cohen 1960). A Cohen's Kappa score k ranges within $[-1, 1]$. A positive k indicates that two observers agree with each other by the degree of k, whereas on the contrary, observers disagree with each other. In the case of classification, suppose that the actual classes of instances are taken as one observer while the predicted classes of them are another observer. Thus, the calculation of agreement between these two observers can be considered as the performance measurement of the model. Cohen's Kappa has a close relationship with ROC but it is more efficient to be calculated than ROC (Ben-David 2008). Cohen's Kappa can be imagined simpler as the performance normalised by its own distribution baselines. This is more meaningful in our case since it enables the comparison between two predictive models from different distributional dataset as they have been normalised respectively. Cohen's Kappa comprises of two parameters: p_o and p_e, where p_o is the accuracy and p_e is the random guess baseline based on its data distribution. The random guess baseline p_e normalise the accuracy p_o on imbalanced dataset. The Equation 2 shows how it is calculated, where ap, pp, an and pn stand for 'actual positive proportion', 'predicted positive proportion', 'actual negative proportion' and 'predicted negative proportion'. The tp, fn fp and tn here stand for the number of 'true positives', 'false negatives', 'false positives' and 'true negatives' respectively.

Case Studies

In this study, experiments were conducted across three commercial games respectively. For each of them, experiments will be performed for predicting the original churn, the original churn with random sampling and the disengagement over varying dates. In each experiment, both game specific features (Runge et al. 2014) and event frequency based data representation (Xie et al. 2014; 2015) will be used to represent the data space respectively. Algorithms such as Decision Tree, Logistic Regression and SVM were applied for all experiments. These algorithms were selected as they have been widely used in this research area (Runge et al. 2014; Borbora et al. 2011; Kawale, Pal, and Srivastava 2009; Borbora and Srivastava 2012).The objective of experiments was to show: firstly, how the classifiers' performances will be affected by the biased dataset that was generated by the original churn, and then, whether the disengagement over varying dates could achieve better performance than the widely used random sampling method for balancing class distributions. As mentioned before, all results in this section are measured by Cohen's Kappa and area under ROC averaged from 10 fold cross validation.

Race Team Manager

The research of this work was firstly inspired by the highly biased issue found in the dataset from this game. While applying the original Churn definition, this dataset shows a highly imbalanced class distribution where the ratio between the churn class (206 players) and non-churn class (54 players) is around 4/1. After applying disengagement over varying dates definition was applied following Equation 1, the distribution between the new churn (62355 players) and new non-churn (51517 players) class became now 1.21/1. For comparison, we also performed random sampling for this data set and get an exact ratio of 1/1 by randomly withdrawing samples from the majority class.

Table 1 shows the results from Cohen's Kappa and ROC tests conducted on the Race Team Manager. In the picture, LR, DT, SVM, RAN stand for Logistic Regression, Decision Tree, Support Vector Machine and Random Guess respectively. S-Features and EF-Features are short for Specific Features and Event Frequency Features. Errors shown are SEMs (standard error of the means) and bold numbers in the table indicate they are significantly better performances according to t-test ($P<0.01$).

As can be seen, performances from both the original churn and its random sampling version are similar to each other. There are nine out of twelve cases where disengagement over varying dates brought significantly better results than others. At the same time, there are only two cases where the original definitions did better. Both of them happened when ROC was used as the measurement and specific features were used for representing the data space. This probably suggests that disengagement over varying dates works better with event frequency based data representation. There is also one case where there is no signification difference among the three definitions. Results from Race Team Manager suggest that predictions achieved better performance for predicting the more balanced dataset created by disengagement over varying than imbalanced classes or ones that were balanced by random withdrawing sampling.

I Am Playr

As mentioned before, the dataset of I Am Playr was simulated to be imbalanced, so that it is consistent with other experiments. After simulation by withdrawing data, the class distribution becomes 4/1 under the original churn definition, which is the same as Race Team Manager. The number of players in churn and non-churn classes become 132 and 33 respectively. While the disengagement over varying dates was applied for balancing, the ratio between disengagement (82 players) and non-disengagement (83 players) was balanced to around 1/1. Same as before, with random withdrawing sampling, the class ratio was exactly 1/1. Note that the number of samples is relatively small (compared with Race Team Manager) because the bias simulation is done after being processed by the filters of the original churn which only focuses on active and high valued players.

Table 2 shows the results from Cohen's Kappa and ROC tests conducted on the dataset of I Am Playr. All notations in this table are the same as the results of Race Team Manager. As it indicates, except for the only case Logistic Regression with specific feature and measure by Kappa, predictions of disengagement over varying dates are significantly better than any other definitions across all cases. Even for the only exception, the p-value from its t-test is 0.0389, which is also significant with $p<0.05$. Thus, same as Race Team Manager, it suggests that classes labelled by the definition of disengagement over varying dates are leading algorithms to the better-balanced prediction performance. At the same time, it is better than the one which uses random withdrawing sampling for balancing.

Lyroke

Likewise, the resultant class distribution from dataset of Lyroke was not highly biased (127/103) when the original churn was applied. We performed the same biasing simulation methods as used on the dataset of I Am Playr to make the ratio of its two classes to be 4/1. After simulation, the number of players in churn class and non-churn class under the original churn definition become 127 and 31 respectively. Then, while the disengagement over varying dates was applied for balancing, the ratio between disengagement (77 players) and non-disengagement (80 players) became nearly 1/1. Same as before, with random withdrawing sampling, the class ratio was exactly 1/1. Similar to I Am Playr, due to the bias simulation being from the filters of the original churn, the number of samples is relatively small.

Table 3 shows the results from Cohen's Kappa and ROC tests conducted on the dataset of Lyroke. Notations in this table are the same as previous ones. Unlike the other two games, three definitions' performances from this game are quite competitive with each other. As can be seen, there is no significant different among three definitions except for two positive cases where disengagement over varying dates performs significantly better when the Logistic Regression and SVM classifiers were used with event frequency based data representation and measured by ROC. The results of this

Table 1: Performance Comparison Among Three Disengagement Definitions on the dataset of Race Team Manager

			Original Churn	Original Churn With Random Sampling	Disengagement Over Varying Dates
Kappa	LR	S-Features	0.08±0.059	0.08±0.059	0.04±0.002
		EF-Features	0.07±0.060	0.07±0.060	**0.80±0.002**
	DT	S-Features	0.22±0.035	0.24±0.043	**0.70±0.002**
		EF-Features	0.14±0.062	0.09±0.043	**0.77±0.002**
	SVM	S-Features	-0.01±0.007	-0.01±0.007	**0.03±0.001**
		EF-Features	0.03±0.067	0.03±0.067	**0.80±0.002**
	RAN	Any	0.00	0.00	0.00
ROC	LR	S-Features	**0.74±0.034**	**0.74±0.034**	0.59±0.002
		EF-Features	0.49±0.031	0.49±0.031	**0.97±0.000**
	DT	S-Features	0.61±0.020	0.62±0.024	**0.87±0.001**
		EF-Features	0.57±0.030	0.55±0.024	**0.89±0.002**
	SVM	S-Features	**0.70±0.031**	**0.70±0.030**	0.42±0.002
		EF-Features	0.51±0.042	0.49±0.042	**0.97±0.001**
	RAN	Any	0.49	0.50	0.50

Table 2: Performance Comparison Among Three Disengagement Definitions on the dataset of I Am Playr

			Original Churn	Original Churn With Random Sampling	Disengagement Over Varying Dates
Kappa	LR	S-Features	0.16±0.094	0.16±0.094	0.39±0.010
		EF-Features	0.04±0.074	0.04±0.074	**0.54±0.012**
	DT	S-Features	0.03±0.071	0.07±0.102	**0.48±0.009**
		EF-Features	0.13±0.071	0.14±0.083	**0.56±0.006**
	SVM	S-Features	-0.01±0.009	-0.01±0.009	**0.29±0.008**
		EF-Features	0.00±0.078	0.00±0.078	**0.57±0.009**
	RAN	Any	0.00	0.00	0.00
ROC	LR	S-Features	0.71±0.055	0.71±0.055	**0.95±0.001**
		EF-Features	0.41±0.052	0.41±0.052	**0.87±0.014**
	DT	S-Features	0.51±0.034	0.53±0.054	**0.74±0.005**
		EF-Features	0.58±0.040	0.58±0.045	**0.78±0.004**
	SVM	S-Features	0.64±0.066	0.70±0.049	**0.94±0.002**
		EF-Features	0.67±0.048	0.46±0.070	**0.94±0.003**
	RAN	Any	0.51	0.50	0.55

Table 3: Performance Comparison Among Three Disengagement Definitions on Lyroke

			Original Churn	Original Churn With Random Sampling	Disengagement Over Varying Dates
Kappa	LR	S-Features	0.42±0.088	0.42±0.088	0.18±0.005
		EF-Features	0.37±0.078	0.37±0.078	0.58±0.005
	DT	S-Features	0.31±0.084	0.28±0.091	0.27±0.003
		EF-Features	0.51±0.089	0.57±0.084	0.53±0.005
	SVM	S-Features	0.15±0.075	0.28±0.103	0.12±0.009
		EF-Features	0.30±0.089	0.30±0.089	0.50±0.024
	RAN	Any	0.00	0.00	0.00
ROC	LR	S-Features	0.91±0.023	0.91±0.023	0.89±0.003
		EF-Features	0.71±0.061	0.71±0.061	**0.99±0.000**
	DT	S-Features	0.66±0.041	0.64±0.048	0.63±0.002
		EF-Features	0.75±0.047	0.78±0.047	0.77±0.003
	SVM	S-Features	0.79±0.027	0.82±0.031	0.88±0.007
		EF-Features	0.69±0.065	0.69±0.065	**0.97±0.004**
	RAN	Any	0.50	0.50	0.56

game suggest that the disengagement over varying dates can at least achieve similar performance after balancing the class distributions and in none of our experiments worse than the random withdrawing sampling method.

Conclusion

Prior works have provided several approaches to predicting player behaviour trends including disengagement and purchasing(Runge et al. 2014; Hadiji et al. 2014; Xie et al. 2014; 2015). Many of them are already able to provide promising performance. However, most labelling methods applied in these works were over restrictive, i.e., the instances in dataset were split into classes by satisfying some specific conditions. Due to this, resultant class distributions of these definitions are often imbalanced. This type of issue can easily lead to biased classifiers during training process. Thus, as consequences, the resultant predictive models will

predict most new incoming examples to the majority side.

To solve the problem, this work introduced a new behaviour trend definition called 'trend over varying dates' that is able to maintain an approximately balanced distribution of resultant classes without losing any samples. By taking disengagement as a typical example, this work explained how the definition can be fitted into this specific predictive task. In a disengagement prediction, rather than selecting disengaging players by hard coded specific conditions, our method tried to partition the database by drawing a flexible 'date line' across the whole dataset. Instead of using a linear one, this 'date line' is formed by 'splitting date segments' from individual players according to their activities respectively. These splitting segments are controlled and generated from two constant parameters prr and por. A player was said to be disengaging or not by Equation 1. In order to get the smallest distance between the resultant classes, a searching algorithm is needed for optimising the two parameters. In this work we applied genetic algorithms for optimising them, but other methods (e.g., gradient search) could be used, too. To evaluate the new definition, we applied it to three different commercial games for balancing the classes before training the predictive models. These three games are all in different genres and developed by two game companies. Both Cohen's kappa and the area under ROC were used as the measurements of the prediction performance. In all three games, the disengagement over varying dates method successfully balanced the distribution of classes (between disengagement and non-disengagement) from 4/1 to nearly 1/1.

There are two main conclusions that can be addressed with the results. Firstly, it is suggested that a biased dataset without balancing can easily lead to failed classifiers as the predictive performance based on the original definition is significant worse than the results from datasets labelled by disengagement over varying dates in almost all cases (except for only two case in Race Team Manager). Secondly, for balancing datasets, the disengagement over varying dates labelling approach can achieve better performance than another widely used random sampling method in most cases. Additionally, except for labelling instances, we also explained that the optimised parameter prr can be used as an indicator of game health and por shows how long does it allow companies to intermit players' disengagement.

The present work focused on applying the concept 'trend over varying dates' to a disengagement prediction task and achieved promising results in most cases. Additionally, further research works will expect to apply the same methodology to more different biased predictive purposes, for instance, purchasing behaviours. Finally, the parameters used by algorithms in experiments of this research are defaults ones at the moment, future works may try to optimise them in order to achieve better performance.

Acknowledgments

The authors would like to thank Bigbit Ltd for supplying data sources of Race Team Manager and WeR Interactive for supplying data sources of both I Am Playr and Lyroke.

References

Alpaydin, E. 2004. *Introduction to Machine Learning*. Adaptive computation and machine learning. MIT Press.

Alpaydin, E. 2010. *Introduction to machine learning*. Adaptive computation and machine learning. MIT Press.

Apt, C., and Weiss, S. 1997. Data mining with decision trees and decision rules. *Future Generation Computer Systems* 13(23):197 – 210. Data Mining.

Ben-David, A. 2008. About the relationship between roc curves and cohen's kappa. *Engineering Applications of Artificial Intelligence* 21(6):874–882.

Borbora, Z. H., and Srivastava, J. 2012. User behavior modelling approach for churn prediction in online games. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, 51–60.

Borbora, Z.; Srivastava, J.; Hsu, K. W.; and Williams, D. 2011. Churn prediction in mmorpgs using player motivation theories and an ensemble approach. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, 157–164.

Campbell, C., and Ying, Y. 2011. *Learning with Support Vector Machines*. Synthesis lectures on artificial intelligence and machine learning. Morgan & Claypool.

Chawla, N. V. 2005. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*. Springer. 853–867.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Davis, J., and Goadrich, M. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240. ACM.

El-Nasr, M.; Drachen, A.; and Canossa, A. 2013. *Game Analytics: Maximizing the Value of Player Data*. Springer.

Hadiji, F.; Sifa, R.; Drachen, A.; Thurau, C.; Kersting, K.; and Bauckhage, C. 2014. Predicting player churn in the wild. In *Proceedings of the Conference on Computational Intelligence and Games (CIG)*.

Hosmer, D., and Lemeshow, S. 2004. *Applied Logistic Regression*. Applied Logistic Regression. Wiley.

Jeni, L. A.; Cohn, J. F.; and De La Torre, F. 2013. Facing imbalanced data–recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, 245–251. IEEE.

Kawale, J.; Pal, A.; and Srivastava, J. 2009. Churn prediction in mmorpgs: A social influence based approach. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, 423–428. IEEE.

Mahlmann, T.; Drachen, A.; Togelius, J.; Canossa, A.; and Yannakakis, G. 2010. Predicting player behavior in tomb

raider: Underworld. In *Computational Intelligence and Games (CIG), 2010 IEEE Symposium on*, 178–185.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2012. *Foundations of Machine Learning*. The MIT Press.

Powers, D. M. W. 2011. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies* 2(1):37–63.

Runge, J.; Gao, P.; Garcin, F.; and Faltings, B. 2014. Churn prediction for high-value players in casual social games. In *2014 IEEE Conference on Computational Intelligence and Games*, 1–8.

Weber, B. G.; John, M.; Mateas, M.; and Jhala, A. 2011. Modeling player retention in madden nfl 11. In *Innovative Applications of Artificial Intelligence (IAAI)*. San Francisco, CA: AAAI Press.

Xie, H.; Kudenko, D.; Devlin, S.; and Cowling, P. 2014. Predicting player disengagement in online games. In *Workshop on Computer Games*, 133–149. Springer.

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

Yannakakis, G. N.; Spronck, P.; Loiacono, D.; and André, E. 2013. Player modeling. *Artificial and Computational Intelligence in Games* 6:45–59.

Yannakakis, G. N. 2012. Game AI revisited. In *Proceedings of the 9th conference on Computing Frontiers*, 285–292. ACM.