Deep Learning for Classifying Battlefield 4 Players

Marjolein de Vries and Pieter Spronck(✉)

Tilburg Center for Cognition and Communication,
Tilburg University, Tilburg, Netherlands
mdv@marjoleindevries.com, p.spronck@tilburguniversity.edu

Abstract. In our research, we aim to predict attributes of human players based on observations of their gameplay. If such predictions can be made with sufficient accuracy, games can use them to automatically adapt to the player's needs. In previous research, however, no conventional classification techniques have been able to achieve accuracies of sufficient height for this purpose. In the present paper, we aim to find out if deep learning networks can be used to build accurate classifiers for gameplay behaviours. We compare a deep learning network with logistic regression and random forests, to predict the platform used by Battlefield 4 players, their nationality and their gaming culture. We find that deep learning networks provide significantly higher accuracies and superior generalization when compared to the more conventional techniques for some of these tasks.

Keywords: Deep learning · Computer games · Player classification

1 Introduction

In recent years, research into gaming and how people interact with games has received much interest. That interest stems partially from the advent of so-called "serious games", i.e., games used for educational purposes [5]. While most serious games use a "one size fits all" approach to their users, adapting a game to a user's personality, skills, and needs, has the potential to make the games more effective [10]. Developers have thus shown interest in modeling the characteristics and behaviours of game players [11].

Games tend to form a rich environment of interaction, from which much knowledge about a player can be gleaned. Previous research has, for instance, focused on modeling a player's personality [3,7], demographics [6], and national culture [1].

While it is relatively easy to gather data on players and their in-game behaviour, attaching meaning to this data is problematic as only player actions can be observed, and not the motivation behind those actions. As a consequence, deriving higher-level interpretations of the observations that allow the game to actually draw conclusions on the player, so that it can adapt effectively to the

R. Poppe et al. (Eds.): INTETAIN 2016, LNICST 178, pp. 161–166, 2017.
DOI: 10.1007/978-3-319-49616-0_15

player's needs, is a tough challenge. Basically, a model is required that is highly accurate in classifying aspects of a player's characteristics. Previous research has shown that player models can be constructed using regular classification techniques, but that such models do not make predictions of sufficient accuracy.

Recently, a resurgence of interest in neural network research has occurred, driven by the increase in computational power and the high availability of data. "Deep learning networks", i.e., neural networks with dozens of layers of high numbers of neural nodes, can be trained to perform classification tasks that have not been handled successfully before. Deep learning has demonstrated its power in classifying images [9], general gameplaying [4] and in recognizing patterns in challenging board games [8].

Our research is driven by the question whether deep learning networks can also be used to classify attributes of game players from observations of their game-play behaviour. The present paper describes our initial research in this respect, where we train a deep learning network with observations of around 100,000 Battlefield 4 players, in an attempt to classify their gaming platform, their nationality and their gaming culture. We chose gaming platform as target variable as it is known for all players, only a limited number of platforms is possible, and because we felt that gaming platform might have subtle interactions with other player features. Nationality was chosen, as previous research [1] has also shown a relation between nationality and gaming behaviour. Moreover, we created a third target variable by dividing the nationalities over five clusters, to reduce the number of target values compared to nationality on its own. We compare the accuracy of our results with the accuracies achieved using logistic regression and random forests.

2 Data

We built a data set of gameplay behaviours of Battlefield 4 players, as such data can be derived easily online. We used a web crawler to acquire the data of about 100,000 Battlefield 4 players from the website *www.bf4stats.com.* As Battlefield 4 can be played on five different platforms (PC, PS3, PS4, Xbox 360, and Xbox One), we made sure to get data for each of those platforms for about 20,000 players.

The web crawler started by retrieving names from the leaderboards, thus making sure that we were mainly gathering data from players who were actually involved with the game. Then we retrieved information on each of those playernames, consisting of statistics such as kill/death ratio, objective scores, scores for different game modes, scores for different roles within the game, and the usage of different weapons. Retrieved statistics concerned the performance of a player up until the moment of retrieval. All player features which are time-dependent were divided by the total playing time. Finally, the dataset was centered and scaled with the R package `caret`, i.e. the mean of every feature was removed to have an average mean of zero and every feature was divided by its standard deviation. During the preprocessing several records were removed, for reasons such as a player having zero playtime, or a playername simply being inaccessible. The resulting dataset contained 99, 912 training examples and 159 features. All features were checked to not have a strong correlation with the target variable platform, which they did not. The distribution of the training examples among the five platforms is shown in Table 1.

Table 1. Platform distribution of the dataset

Platform	PC	PS3	PS4	X360	XOne
Nr of examples	19,839	20,007	20,010	20,021	20,035
Percentage (%)	19.86	20.02	20.03	20.04	20.05

Nationality was known for only 10, 665 out of the 99, 912 players. 155 different nationalities were present in the dataset. We selected only countries with 50 or more representatives, resulting in a dataset consisting of 9, 770 training examples with 33 different nationalities. The most common nationality was American, with 1, 897 players. k-Means clustering with $k = 5$ was performed on the averages for each country, resulting in the five different clusters as shown in Table 2. An ANOVA on the effect of cluster on each of the six Hofstede dimensions [2] was performed. Cluster 2 was left out in this analysis, as it consisted of only one nationality. A significant effect at $p < 0.05$ level of cluster on the dimensions Power Distance ($F(3, 28) = 9.34, p < 0.001$), Individualism ($F(3, 28) = 15.25, p < 0.001$), Long Term Orientation ($F(3, 28) = 4.28, p = 0.013$) and Indulgence ($F(3, 28) = 8.14, p < 0.001$) was found. A significant effect was found between every pair of clusters, except for cluster 3 and 5. As Hofstede's cultural dimensions are clearly distinguishing the clusters we found, we may assume that the clusters can be regarded as a representation of culture.

Table 2. k-Means clusters with $k = 5$

Nr	Size	Members
1	934	China, Chech Republic, Russia, South Korea, Turkey, Ukraine
2	101	Saudi Arabia
3	3612	Australia, Belgium, Britain, Canada, France, Poland, USA
4	1071	Argentina, Brazil, Chile, Mexico
5	4052	Austria, Denmark, Finland, Germany, Italy, Japan, the Netherlands, New Zealand, Norway, Portugal, Spain, South Africa, Sweden, Switzerland

3 Data Analysis Models

We used three different data analysis models: (1) logistic regression, (2) random forest, and (3) deep learning network.

Multinomial Logistic regression with regularization is a widely used model for data analysis. It is an adaption of the Linear Regression model in order to make it usable for classification. There are different methods for subset selection for this model, for which the standard is the shrinkage method, by which a penalty is given on large weights in order to prevent overfitting. We used the R package `glmnet` for logistic regression.

A Random Forest is a collection of decision trees, where the output of all decision trees is combined into one classification label. We used the R package `randomForest` to implement this classifier.

A deep learning network is a neural network that may have many layers, and a great many nodes per layer. For training the network we use standard backpropagation, using the Python `Keras` library, which is built onto the `Theano` library.

Each model has different parameters which need to be optimized. To do so, we divided the dataset into a training, validation, and test set before learning the models for predicting platform. The training set consists of $60,000$ examples, the validation set consists of $20,000$ examples and the test set consists of the remaining $19,912$ examples. The different models are trained on the training set and their error percentage is evaluated on the validation set in order to determine the best parameter(s) for each model. The test set is used to determine the final accuracy of the model. Because of the smaller size of the dataset for nationality and cluster culture, we used 10-fold cross-validation instead of a seperate training and validation set. The cross-validation set for this approach consists of $8,000$ examples, and the test consists of the remaining $1,770$ examples. All models use the same division of training examples among the sets for each type of target variable.

4 Results

The results for all four models are shown in Table 3.

For the logistic regression, we used different values for α and λ.

For the random forest, we tested different numbers of trees (ranging from 10 to 5,000) and different numbers of minimum leaf node sizes (ranging from 1 to 500). In total, we compared 56 different configurations. However, these results are in all cases worse than those for logistic regression.

For the deep learning network, we considered a perceptron network with one hidden layer and with two hidden layers, and for classifying platform also with three hidden layers, with different numbers of hidden nodes in each layer. The number of hidden units varied within the values $(128, 256, 512, 1024)$. Moreover, the values $(0.01, 0.001, 0.0001)$ were considered for the learning rate. As the outcome of a neural network is dependent on the start weights, we built each model five times with different random start weights. The final error percentage for a set of parameters was calculated by averaging over these five networks.

The error percentage on validation set for classifying platform was 18.99 for one hidden layer, 17.69 for two hidden layers and 17.24 for three hidden layers.

The best performing neural network for classifying platform has three layers with 1024 hidden nodes each. Due to the diminishing returns of adding more layers, we did not test what adding a fourth layer would do.

The table shows that neural networks show a significant better performance than other models when classifying platform. For classifying nationality and cluster culture, they show similar performance as logistic regression. This is not surprising, as the dataset used for nationality and cluster culture consisted of only 10 % of the examples for platform, and neural networks perform better on large datasets.

From the table it can also be seen that the results for almost all models are slightly better on the test set than on the validation set. This demonstrates that they all generalize well.

An analysis of the confusion tables for the neural networks on platform showed that identifying a PC player is most easy. Xbox One players are sometimes confused with PS4 players, and Xbox 360 players are sometimes confused with PS3 players. Evidently, the generation of the console has more impact on playstyle than the type of console.

Table 3. Error percentage on validation and test sets

Model	Platform		Nationality		Cluster	
	Validation	Test	Validation	Test	Validation	Test
Baseline	79.49	79.79	80.48	81.07	58.64	58.02
Logistic regression	20.39	20.33	55.88	55.93	39.46	39.77
Random forest	23.37	22.99	60.24	60.17	43.59	42.77
Neural network	17.24	16.43	56.10	55.75	39.67	39.31

5 Conclusion

Our research goal was to investigate to what extent a deep learning neural network can derive players' characteristics from their gaming behaviour. We remark that a neural network with three layers is a fairly simple deep learning network, but the gain in accuracy between two and three layers in the present setup did not warrant adding more layers. However, even the three-layer neural network has proven to perform significantly better than conventional approaches for classifying platform, with a classification accuracy of 83.57%. On a much smaller dataset, neural networks perform similar to logistic regression to classify nationality and culture, with a classification accuracy of 55.75% for nationality and of 39.31% for cluster culture. We are currently investigating whether using a larger dataset improves the neural network performance for this task.

We have shown that neural networks, by their strong performance, are a viable modeling approach to player attribute classification, and thus have the potential to open the road to actual automatic game adaptation based on player observations. However, improved accuracy could definitely be achieved when we use more extensive feature sets. In this research, all the features were snapshots of statistics at a specific time. Previous research has shown that far better results can be achieved by including features that express behaviour change over time [6]. Naturally, we also need to investigate predicting different player attributes.

Note: A white paper with many details on the experiments described in this paper is available from the authors on request. The dataset is also available from the same source.

References

1. Bialas, M., Tekofsky, S., Spronck, P.: Cultural influences on play style. In: Proceedings of the 2014 IEEE Conference on Computational Intelligence in Games, pp. 271–277. IEEE Press (2014)
2. Hofstede, G., Hofstede, G.J., Minkov, M.: Cultures and Organizations: Software of the Mind, revised and expanded 3rd edn. McGraw-Hill, New York (2010)
3. Van Lankveld, G., Spronck, P., Van den Herik, J., Arntz, A.: Games as personality profiling tools. In: 2011 IEEE Conference on Computational Intelligence in Games, pp. 197–202. IEEE Press (2011)
4. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015)
5. Prensky, M.: Digital Game-Based Learning. McGraw-Hill, New York (2001)
6. Tekofsky, S., Spronck, P., Goudbeek, M., Plaat, A., Van den Herik, J.: Past our prime: a study of age & play style development in battlefield 3. IEEE Trans. Comput. Intell. AI Games **7**(3), 292–303 (2015)
7. Tekofsky, S., Spronck, P., Plaat, A., Van den Herik, J., Broersen, J.: PsyOps: personality assessment through gaming behavior. In: Proceedings of the 2013 Foundations of Digital Games Conference (2013)
8. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, L., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of Go with deep neural networks and tree search. Nature **529**, 484–489 (2016)
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the ICLR 2015 (2015)
10. Westra, J., Dignum, F.P.M., Dignum, M.V.: Organizing scalable adaptation in serious games. In: Proceedings of the 3rd International Workshop on the uses of Agents for Education, Games and Simulations (2011)
11. Yannakakis, G.N., Spronck, P., Loiacono, D., André, E.: Player modeling. In: Artificial and Computational Intelligence in Games. DFU, vol. 6, pp. 45–59. Dagstuhl, Germany (2013)