Clustering steam user behavior data using K-Prototypes algorithm

Kiefer Stefano Ranti*, Kelvin Salim, and Abba Suganda Girsang

Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

* kiefer.ranti@binus.ac.id

Abstract. The usage of user telemetry to gather player behavioral data on video games can be very beneficial to game developers with a certain business model. With the help of user telemetry in game development, it can provide access to data on user behavior from installed game clients' platform such as Steam. These behavioral data can be used to find out the Steam user behavioral patterns on playtime distributions that can be studied by developers in order to have a deeper understanding of the behaviors of their players. In this study, the data are clustered using the k-prototypes algorithm, a combination of k-means and k-modes algorithm that can be used to cluster mixed attributes. The result shows that the clusters represent the types and preferences of the players.

1. Introduction

With the ever-increasing popularity of gaming to the mainstream population, a lot of game companies such as Valve, EA, and Ubisoft are using telemetry through the installed game client software distribution platform to collect user behavioral data. These data, which usually contains information such as the user purchasing behavior and their interaction with games are then studied by these companies in a form of analytics in order to help them have a clearer understanding the behaviors of the players of their games, in other words, their customer. It helps them to develop, design, and market their product better tailored to their customer's behavior and needs.

The increasing trend of this user behavior analysis is in connection with the rising popularity of the free-to-play (F2P) and freemium business models that are a very popular model to be implemented in many popular games worldwide. These business models require constant monitoring and evaluation of player population behavior in order to drive revenue. The practice has become the norm in the game industry.

There are many studies that have been done on this specific domain of user behavior analysis, mostly using unsupervised clustering method on Steam data since steam is one of the biggest game client platform available right now. Regarding the Steam platform, (1) mined steam player profiles that resulting in descriptive statistics. The study shows a good correlation between sales data from the sample and actual sales data reported by game development companies. (2) has done a large-scale experiment to provide insight into the patterns around playtime in the games bought and played by steam users, as well as patterns about the user themselves. (3) did a study to cluster the player behavioral data from two major commercial game title using two algorithms; SIVM and k-means, resulting in different insight for each algorithm. (4) uses five different algorithms to cluster the behavioral telemetry dataset from World of Warcraft and the result shows that the method chosen has an impact on

the result. (5) shows the importance of differentiating users for freemium games and highlighted that doubling the player base does not also double the revenue.

In this paper, another method will be used to cluster the user behavior telemetry data. The k-prototypes algorithm will be used as it integrates both k-means and k-modes algorithms to cluster objects with mixed numeric and categorical attributes (6). This algorithm is chosen because the dataset used is user telemetry data that contains mixed attributes, both numerical e.g. playtime and categorical e.g. game genre.

By using this algorithm, the steam users' behavior is generated into some scenario cluster. The scenario is the number of clusters is 4, 5 and 6. To generate the dataset behavior of Steam's users, the users are chosen are those who spent 40% or more of their total accumulative playtime. The purpose of choosing this specific data is to find out exactly which games they play most. Although just by looking at the percentage doesn't really guarantee that the player actually plays the game a lot, so a player with below 100 hours of total playtime will be excluded.

2. Related Work

User telemetry is a data gathered through the use of software that contains a user's interaction information. In this study case, from the perspective of the game industry, it is the information gathered from a user that logged in on a game client, e.g. how often does the user log in, how many hours did the user play for this week, and what games did the user bought, all of these are part of the player behavior telemetry (7). Any action that the user takes can be recorded on the installed game client. These recorded data are then converted into matrices that can be used for analytics.

Clustering is a process of grouping a set of objects where into groups of related items (8). Objects in the same cluster have more similarities than those in the other cluster (9). It is a fundamentally important task in machine learning, data visualization, computer vision, and other domains (8). the k-means algorithm is one of the oldest, most popular unsupervised learning methods to use for clustering because it is simple and easy to implement (10,11). But it had a limitation, that is the k-means algorithm can only be used on datasets with numerical attributes, making it unusable on categorical attribute data (12). To solve this problem, (6) proposed the k-modes algorithm, a modification of the k-means algorithm that is faster and usable on categorical attribute data. The difference is instead of distances, k-modes uses dissimilarities to calculate distance. And it uses modes, instead of means (6,13). Although the k-modes is able to cluster categorical data, the algorithm cannot cluster mixed attributes data. The dataset that was used in this study contains both numerical and categorical attributes, which makes it a mixed attributes data. To handle both of these attributes, the k-prototypes algorithm can be a solution (14).

In this paper, the k-prototypes algorithm was used to cluster the data obtained from the steam user profiles. The k-prototypes algorithm is essentially a k-means and k-modes algorithm combined to tackle mixed data that contains both numeric and categorical attributes. The k-prototypes algorithm process of clustering is similar to the k-means algorithm except that it uses the approach of the k-modes algorithm to update categorical attribute values (15,16). Because this algorithm uses the same clustering process as k-means, it still preserves the efficiency of k-means algorithm although it is still slower than the k-modes algorithm as k-prototypes algorithm needed more iteration to converge. The k-prototypes algorithm is more useful for data in real-world cases as encountered objects usually a mix of numeric and categorical objects.

3. Proposed Method

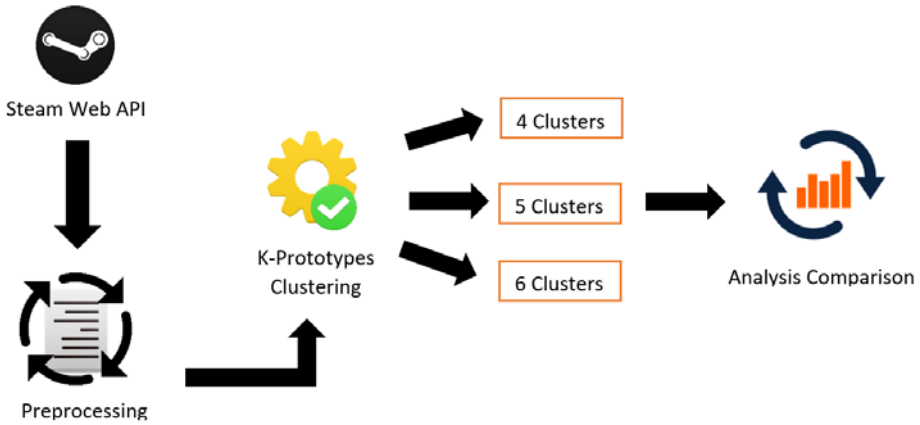

Fig 1. Research steps

Fig. 1 shows the steps that are taken in this study. The first step is fetching the data from steam web API. Then preprocessing is performed on the fetched data by applying the min-max normalization method. The k-prototypes algorithm is applied to the normalized data three times with a different number of maximum clusters for each iteration.The clustering results are then analyzed.

In this paper, a clustering analysis is performed on a 500 player's dataset totaling 3 million hours of playtime. The players are picked randomly from a game review pages, then checked whether their profile is public or not. If the profile is public, then the player baseID are recorded then will be used to get the player information through the Steam web API. The data contains a lot of information about the player which then filtered so only the information needed is taken, such as the player total playtime, total games the player owned or bought, game titles, the game's genre, and developers.

There are two important features needed in this study; the total accumulated playtime for all the games a player own/played and the playtime of a game where the player spent 40% or more of their total accumulative playtime. There are 2 main constraints in the data collection; If the player doesn't play a game for 40% of their total accumulative playtime and if the player doesn't have 100 or more total playtime their profiles are not included in the dataset used for the clustering. These two constraints are made to make sure the player actually more partial the game. The acquired data are parsed to a more readable format, where the columns are labeled as such; "game count", "total playtime", "game title", "game playtime", "genre", "developer", "percentage".

Table 1. Sample data

Game Count	Total Playtime	Game Title	Game Playtime	Genre	Developer	Percentage
3128	14524	Counter-Strike: Global Offensive	7788	Action	Valve	54%

The filtered player dataset is then processed. Any symbols and special characters in the dataset are removed as it can interfere with the clustering process. Numerical data such as "total playtime", "game count", "title playtime", and "percentage" are processed using a feature scaling method, the min-max normalization, that consist of scaling the range of features to scale the range from 0 to 1 (17). It is a technique that can help in improving the k-means clustering algorithm (18,19), and this fact also applies to the k-prototypes algorithm because of the similarity between the two algorithms. Eqt (1) shows the general equation on min-max normalization.

$$x' = \frac{x - min(x)}{max(x) - min(x)} \quad (1)$$

Table 2. Sample data after preprocessing

Game Count	Total Playtime	Game Title	Game Playtime	Genre	Developer	Percentage
0.1403186	0.139244027	Counter strike global offensive	0.124004333	action	valve	0.2437329

After being processed, the data is ready for clustering. Based on the literature (6,14,20), we can conclude that there are 4 steps in implementing the algorithm:

1. **Reading the parameter**. For this step all the parameters from the dataset are being read, such as:
    - n = Number of records
    - k = Maximum number of clusters
    - Each categorical attributes' number of categories
    - Attribute names and types
2. **Initial prototypes selection**. In this step, k objects are selected randomly as the initial prototype for the k cluster.
3. **Initial allocation**. Each object in the dataset is assigned to a cluster that has a minimum difference with its prototype using a dissimilarity measure. After clustering, the prototype is updated after each assignment.
4. **Reallocation**. Now, the prototype from the previous and current cluster will be updated. In the process of executing the algorithm on the python program, the console will be showing "moves". This means that the console is informing us that some objects have changed clusters in the process. When the moves are equals to zero, it means that the algorithm has acquired the best result in this iteration.

4. Analysis Results

With the intention to investigate how the player invests their time on games they play, a cluster analysis is conducted using the k-prototypes algorithm. By clustering this data, the result can provide representation and summarization on how the players are grouped according to their playtimes. The results show how the players spend the majority of their time and various types of player that exists on the Steam platform.

Running the k-prototypes clustering process shows that every cluster contains one most played game title. For example, when clustering the data with k=5, one of the clusters contain players that use the majority of their playtime playing Garry's Mod. Although there are still other games that got included in the cluster. Clustering with k=4 produces a result that is analyzable although with k=5 shows some interesting patterns and separability. Going for k=6, the first five clusters will just split into clusters that focus on each majority's most played games and for the sixth cluster, it will contain the games made by Valve such as CS: GO and Dota 2. The result that will be discussed here is with k=5, and each cluster that was generated will be named accordingly with their content that represent them best. Descriptions for each cluster are as follows:

1. **First cluster**, the Action cluster.Represents 47.2 per cent of the data set. Contain players that play their most played game for 40-50% of their total playtime. Biggest of the five clusters, most of the games are action genre.
2. **Second cluster**, the Small Inventory cluster.Represents 5.1 percent of the data set. Contain players with a high percentage of playtime of their most played games, mostly over 60%. Players in here don't own many games, they own around 30-80 games per player, that pales in comparison to players on other clusters. This cluster will be called Small Inventory Cluster.
3. **Third cluster**, the Simulations cluster.Represents 16.5 of the data set. Contain players with a lot of playtime on simulation games and unusually high total playtime. Majority of games contained are Arma and Garry's mod.
4. **Fourth cluster**, the Competitive cluster.Represents 19.5 per cent of the data set. Contain players that mainly only focusing on playing competitive multiplayer games, around 60 to 90% of their total playtime.Games such as Counter-Strike and Dota 2.
5. **Fifth cluster**, the mixed cluster.Represents 11.5 per cent of the data set. Similar to the first cluster, contain players that play their most played game for 40-50% of their total playtime, except this cluster contains a mixed variety genre of games.

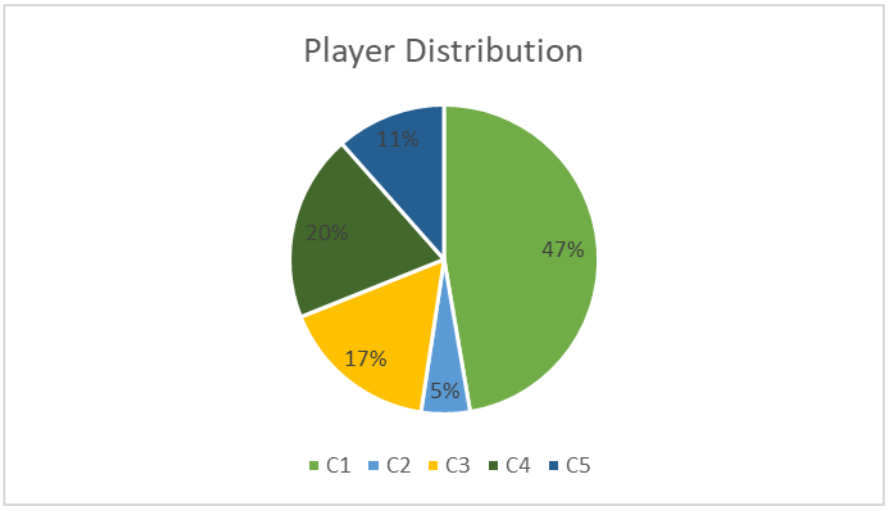

Fig 2. Percentage of players on each cluster

Fig. 2 shows how the clustering process grouped the player data. The players are distributed to each of the five clusters, and the first cluster is where most players get grouped. The first cluster contains players that play action games for 40 percent or more of their playtime. This is not surprising as the majority of the game available either free or for sale on steam are action genre. The second cluster, the small inventory cluster, is where players that owned or bought a small quantity of games are grouped. The players in this cluster seem to focus on playing and owning just a few games. Since they are just playing a few select games, the playtime percentage is quite high (the majority of it exceeds 60%), since the playtime of their most played game doesn't differ much with their total playtime. The third cluster is the simulations cluster. This cluster contains players that mainly play games with simulation genre. The majority of these players has total playtime over 2000 hours and plays one of the most popular games on Steam, Garry's Mod. The high playtime is probably due to the fact that simulation games have high replay ability since most of the games on that genre do not have stories, thus making the game have no concrete ending. The fourth cluster is the competitive cluster. In this cluster, the players seem to play competitive multiplayer games almost exclusively. The players have high playtime percentages. Games the players play here are; Counter-Strike, Dota 2, and Team Fortress 2. All of them are developed and published by Valve, the owner of Steam. The fifth cluster is where the players have a variety of interests. Pretty much similar with the first cluster,except that all other genres in the dataset are also grouped in this cluster.

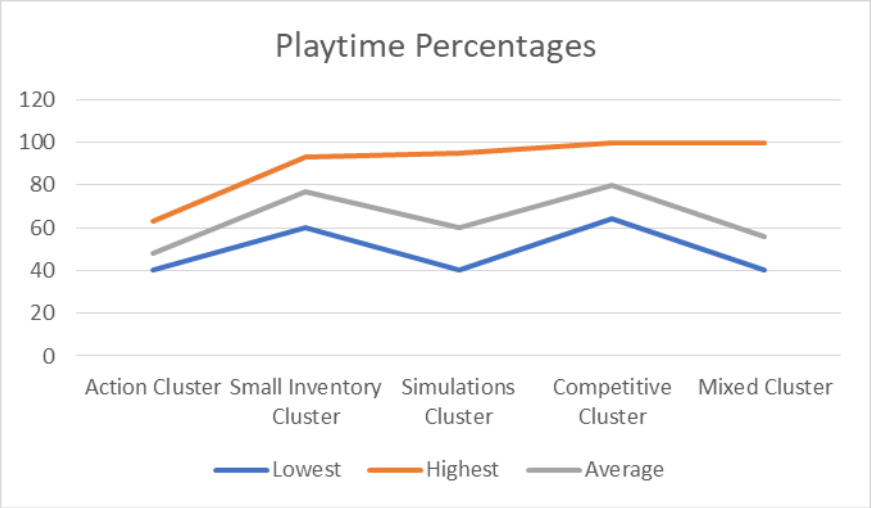

Fig 3. Highest, lowest, and average of player most played game percentage

Fig.3 shows the playtime percentage distribution on each cluster. Lowest line shows the lowest playtime percentage of each cluster, and highest line shows the highest playtime percentage of each cluster. The average percentage of each clusters are 48%, 77%, 60%, 80%, and 56%, respectively. Both cluster 4 and 5 have players that had 100% of their playtime just exclusively playing one game.

5. Conclusion

All the analysis presented in this study is focused on the playtime of the players on the Steam platform, covering 500 players with more than 3 million hours of total playtime and 40+ unique game titles. Since the data contains only players that more partial to one specific game, the clustering results show that gamers are more partial to action games. Almost half of the players on the dataset play action games, which are normal, considering that the majority of the games available on the steam platform are action genre.

Future work should be focusing on even more large-scale analysis, using much larger dataset for even more detailed analysis of player behaviour. Game review features can also be added for more information on games the player plays. There is also the use to make an improvement on Steam's recommendation system. The recommender system can also be used to help developers identify high-value users to prioritize.

References

[1] Orland K. Introducing Steam Gauge: Ars reveals Steam's most popular games [Internet]. Arstechnica.Com. 2014 [cited 2019 Feb 17]. p. 281. Available from: http://arstechnica.com/gaming/2014/04/introducing-steam-gauge-ars-reveals-steams-most-popular-games/%5Cnhttp://arstechnica.com/gaming/2014/04/introducing-steam-gauge-ars-reveals-steams-most-popular-games/2/

[2] Sifa R, Augustin S, Drachen A, Bauckhage C. Large-Scale Cross-Game Player Behavior Analysis on Steam. Elev AAAI Conf Artif Intell Interact Digit Entertain. 2015;198–204.

[3] Drachen A, Sifa R, Bauckhage C, Thurau C. Guns, swords and data: Clustering of

player behavior in computer games in the wild. 2012 IEEE Conf Comput Intell Games, CIG 2012. 2012;163–70.
[4] Drachen A, Thurau C, Bauckhage C. A Comparison of Methods for Player Clustering via Behavioral Telemetry. Proc FDG '13. 2013;245–52.
[5] Lim N. Freemium games are not normal [Internet]. Gamasutra. URL: http://www. gamasutra. com/blogs/ …. 2012 [cited 2019 Feb 17]. Available from: https://scholar.google.com/scholar?start=10&q=allintitle:+freemium&hl=en&as_sdt=0,5#6
[6] Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min Knowl Discov. 1998;2(3):283–304.
[7] Bernhaupt R. Game User Experience Evaluation. Bernhaupt R, editor. Cham: Springer International Publishing; 2015. 1-8 p. (Human–Computer Interaction Series).
[8] Awasthi P, Charikar M, Krishnaswamy R, Sinop AK. The Hardness of Approximation of Euclidean k-means. 2015;1–14.
[9] Jain AK, Lansing E. Data Clustering : 50 Years Beyond K-Means 1 Anil K . Jain Michigan State University. Pattern Recognit Lett. 2011;31(8):651–666.
[10] Slamet C, Rahman A, Ramdhani MA, Dharmalaksana W. Clustering the verses of the holy qur'an using K-means algorithm. Asian J Inf Technol. 2016;15(24):5159–62.
[11] Gan G, Ng MKP. K-Means Clustering With Outlier Removal. Pattern Recognit Lett. 2017;90:8–14.
[12] Jiang F, Liu G, Du J, Sui Y. Initialization of K-modes clustering using outlier detection techniques. Inf Sci (Ny). 2016;332:167–83.
[13] Duan Q, Yang YL, Li Y. Rough K-modes clustering algorithm based on entropy. IAENG Int J Comput Sci. 2017;44(1):13–8.
[14] Sangam RS, Om H. An equi-biased k-prototypes algorithm for clustering mixed-type data. Sadhana - Acad Proc Eng Sci. 2018;43(3):1–12.
[15] Ji J, Bai T, Zhou C, Ma C, Wang Z. An improved k-prototypes clustering algorithm for mixed numeric and categorical data. Neurocomputing. 2013;120:590–6.
[16] Arora P, Deepali, Varshney S. Analysis of K-Means and K-Medoids Algorithm for Big Data. Phys Procedia. 2016;78(December 2015):507–12.
[17] Patro SGK, sahu KK. Normalization: A Preprocessing Stage. Iarjset. 2015;20–2.
[18] Choudhary A, Sharma P, Singh M. Improving K-means through better initialization and normalization. 2016 Int Conf Adv Comput Commun Informatics, ICACCI 2016. 2016;2415–9.
[19] Eesa AS, Arabo WK. A Normalization Methods for Backpropagation: A Comparative Study. Sci J Univ Zakho. 2017;5(4):319.
[20] Guo D, Chen Y, Chen J. A K-Prototypes Algorithm Based on Adaptive Determination of the Initial Centroids. 1(2):116–21.