Cluster Based Analysis on Video Game Sales

Katie Christensen
Computer Science Department
Western Washington University
Bellingham, WA
chris90@wwu.edu

Ian Cullum
Computer Science Department
Western Washington University
Bellingham, WA
cullumi@wwu.edu

Seth Walloch
Computer Science Department
Western Washington University
Bellingham, WA
wallocs@wwu.edu

Abstract

Essential to data mining are learning tasks that can handle high-dimensional data in any form. Visualizing, interpreting, and analyzing data that is unstructured is a job for clustering. Clustering provides information about the nature of, and patterns within, a dataset that wouldn't have been possible to understand without organizing it into clusters. Clustering is a fundamental machine learning technique, and algorithms that are centroid-based, density-based, hierarchical, and many others, are extensively utilized in the data mining field.

This paper walks through an application of clustering to the video game industry. Individual video games are organized into multiple clusters based on their sales information from three regions around the world.

Keywords – machine learning, unsupervised, clustering, partitioning-based, hierarchical-based, data mining

1. Introduction

The video game market extensively holds a large amount of data, specifically on sales information from around the world. Taking this data and analyzing it is a difficult task without popular data mining learning tasks. The use of clustering to inspect large, unstructured data on video game sales is explained thoroughly within this paper. The goal of the investigation was to find natural groupings within the sales records, and cluster the games contained within the dataset into said clusters.

The dataset itself contained an abundant amount of information ranging from numerical sales in different regions of the world, to categorical features such as the studio that published each game. By taking data solely on video game sales, it's anticipated to be able to make further predictions about the video game industry as a whole after the completion of this analysis.

2. Background

The dataset analyzed was titled "Video Game Sales," and can be found on Kaggle.com. The dataset itself had eleven different attributes, consisting of the name of the video game, the platform it was released on, the year and genre the game was released under, the publisher of the game, and its sales information. The sales data was broken into five columns which separated the sales by region. The regions were NA (North America), EU (Europe), JP (Japan), and Other. The "Other_Sales" attribute was anywhere else in the world that the game was sold. The last attribute that was a part of the sales section of the dataset was "Global_Sales" which combined all of the sales in each region under a simple view of total sales.

The first four sales columns were used as the input data into the unsupervised clustering learning model, explained more in depth later in the report. The remaining data was then analyzed and explored for trends or patterns around the resulting clusters. This was a unique approach to this kind of problem because many of the similar works had solely focused on the sales columns and not on the categorical features.

The findings for this project were interesting when comparing all the different sales attributes with the different genres, years, and publishers. It displayed multiple different patterns that weren't previously anticipated to be discovered.

3. Related Work

As previously stated, the original dataset used for this project originated from Kaggle.com. Many similar exploratory efforts were previously conducted using this same dataset and those that are similar, and many of these can also be found on Kaggle's website.

4. Implementation

The environment used was a "Google Colaboratory" notebook completed in Python. This provided a clear and organized structure of the executed methods.

a) Exploratory Analysis:

The procedure began with a meticulous investigation of the dataset. It was pivotal to know the number of rows, columns, and data points contained within the dataset as machine learning models require an abundant amount of data. It was found that the dataset contained over 180,000 data points which was plentiful for the scale of this study.

A seaborn displot was then created to visualize the columns containing missing values, if any existed. From the resulting graph, it was clear that the dataset did contain missing values, so handling these would be a necessary future step. In order to know if imputation of the missing values was required, the number of rows containing missing values was then calculated.

Finally, it was necessary to check the datatypes of each of the attributes since these types of machine learning models can only take in numerical data as input.

b) Data Preprocessing / Feature Extraction:

From the exploration explained above, it was decided that missing values were to be dropped instead of imputed because they only made up 1.8% of the rows in the dataset; the study could afford to lose these rows. The dataset was then shuffled to avoid the model learning any potential patterns that might have been organized into the dataset.

Specific columns of the dataset were chosen to become input into the clustering models. These included all the sales information features, "NA_Sales," "EU_Sales," "JP_Sales," and "Other_Sales," excluding "Global_Sales." This column was left out because it was intended to compare the resulting clusters against this feature as a visualization technique during performance evaluation, explained more in the following section. These columns were converted into Numpy nd-arrays, and they were ready for model training.

c) Models Developed:

Partitioning and hierarchical-based algorithms were selected to evaluate. These included K-means, MeanShift, and Agglomerative clustering, where the number of clusters was set to 3 as the parameter for the two centroid-based algorithms.

4. Evaluation

a) Evaluation Metric - Silhouette Coefficient

The chosen evaluation metric was The Silhouette Coefficient. This metric is a value defined for each data point that is composed of two scores, (1) the mean distance between a sample and all other samples in the same cluster, and (2) the mean distance between a sample and all samples in the next nearest cluster. The resulting coefficient is bounded by [-1,1] where values close to -1 indicate an incorrect clustering, values close to 0 indicate overlapping clusters, and values close to 1 indicate correct, accurate clustering. This allowed for a clear understanding that clustering techniques with a score closer to 1 are more effective than techniques with a lower score. The Silhouette Coefficient is a respective and popular evaluation metric for clustering implementations.

b) Clustering Approaches

Three clustering algorithms were tested and compared in order to determine the optimal model for analyzing the chosen dataset. These algorithms included K-Means, Agglomerative, and Mean Shift.

i. K-Means Clustering

K-Means clustering is a centroid-based algorithm that clusters data by separating samples into n groups of equal variance, and minimizes the sum-of-squares within each cluster. The K-Means algorithm, shown in **Figure 1**, gave the following results with a Silhouette Score of 0.8436.

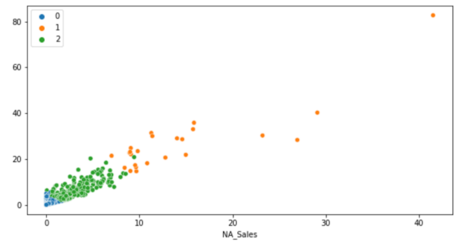

Figure 1. K-Means clustering, NA_Sales vs Global_Sales

ii. Agglomerative Clustering

Agglomerative clustering is a hierarchical strategy useful for handling a very large number of data. Since the dataset contains over 16,000 instances this algorithm seemed appropriate to test, shown in **Figure 2**. However, it resulted in a Silhouette Score of 0.8524, only marginally better than K-Means.

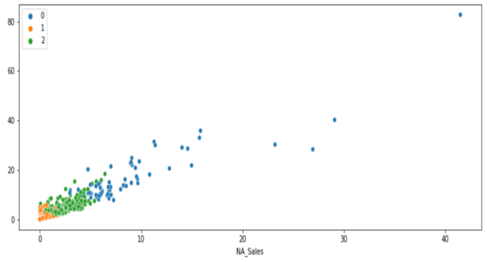

Figure 2. Agglomerative clustering, NA_Sales vs Global_Sales

iii. Mean Shift Clustering

Mean Shift clustering is an approach that aims to discover "blobs" in smooth density samples. It is another centroid-based algorithm, one that calculates mean shift vectors that point in the direction of the densest area in the region, where the centroid gets updated to. This process repeats until a certain threshold is met.

The resulting Silhouette Score was 0.9567, and shown in **Figure 3**, the value that appears to be an outlier was actually placed into a separate cluster. One can conclude that this is the reason the algorithm proved to be most successful.

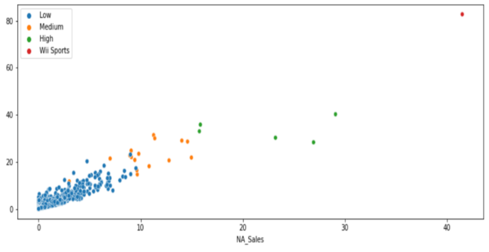

Figure 3. MeanShift clustering, NA_Sales vs Global_Sales

c) Cluster Analysis

After MeanShift clustering was decided upon as the ideal approach to focus on, the dataset was plotted out (highlighted by the clusters generated by MeanShift clustering) in several different ways with the goal of highlighting unique observations and trends in the data. Specifically, the data was graphed by Platform, Genre, and by Year. In addition to displaying the data points by global sales and the aforementioned categories, the graphs generated also displayed a line chart. This feature of the graph communicates the count of games that sold over 100,000 copies for each value. For the purposes of analysis, the term “successful game” will be defined as a game that sold at least 100,000 copies by the time this dataset was created (the dataset only contains games with 100,000 or more sales in or prior to the year 2019).

i. Platform

Analyzing **Figure 4**, the Platform chart, yielded primarily a correlation between the number of successful titles and the number of games that fall under higher sales clusters. Specifically, platforms with a greater number of successful games are more likely to have games that fall under higher sales clusters. In other words, platforms with lots of games that sell well also tend to support games that sell better overall. It's not immediately clear the source of this correlation, though a possible, speculative explanation could be that more popular platforms sell more copies and therefore more games are made for those platforms leading to a higher number of successful games on that platform. Essentially, a feedback loop.

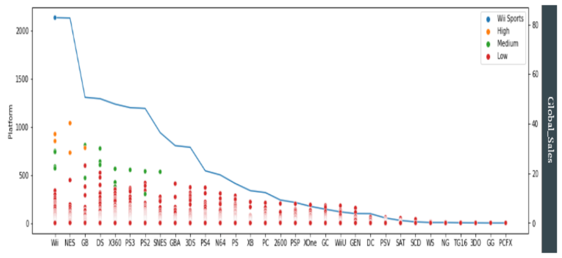

Figure 4. Platform vs count (left-axis) and Global_Sales (right-axis)

ii. Genre

Analysis of the Genre chart, **Figure 5**, showed very similar results, however, this correlation was less prominent compared to Platforms.

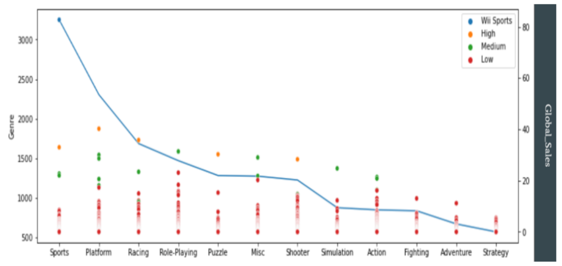

Figure 5. Genre vs count (left-axis) and Global_Sales (right-axis)

iii. Year

Comparing our cluster data to Year produced the most amount of information, especially when paired with rough knowledge of real life events. In **Figure 5**, there are several pieces of information worth highlighting.

The first being that generally, games with earlier release dates have had more time to sell copies. Another point to consider is that games with later release dates will likely sell more copies in the future. It’s also important to note that the number of successful games can likely be expected to increase over time along the roughly exponential rate observed. Next, there are a few games released in earlier years that fall into higher sales clusters that are clear outliers for their given years. Finally, games that have kept pace with games released much later may very well keep pace with newer sales numbers, likewise some more recent titles can likely be expected to keep rising in sales for years to come.

It is worth noting that years that are believed to have been very successful for the games industry are not included on this timeline, namely 2020 and onward, when worldwide quarantines took effect. Outliers with earlier release dates could likely represent titles which have continued to produce and sell copies longer than most of their peers with the same release dates, which could account for the sharp increase in sales they show comparatively. It is also likely that the overall curve shown correlates to the overall success of the games industry.

A reasonable characterization of this data would be that the overall ceiling for game sales has increased along this curve. Naturally, the floor of this graph does not include games that sold less than 100,000 copies, so it is not clear how many games are made overall for any given year or what percentage of games overall make at least 100,000 in sales.

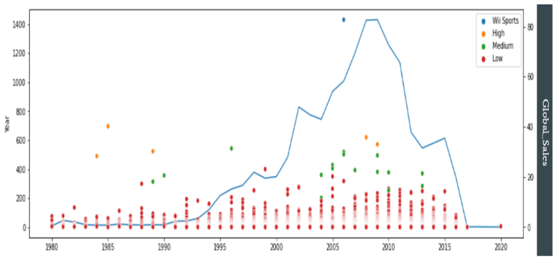

Figure 6. Year vs count (left-axis) and Global_Sales (right-axis)

5. Conclusions and Future Work

Now that three clusters have been created and labeled, these can then be used as classes in a future classification supervised learning task. The goal would be to receive a video game with categorical information as input, such as genre, platform, publisher, etc., and classify this game into one of the three clusters: labeling it with either a “low,” “medium,” or “high” class label. This would provide key information to, for example, game developers because one would be able to input a certain platform and they would be told if their game is going to sell high or low compared to a different platform.

In conclusion, clustering analysis provides a lot of insights into unstructured data. It’s evidently an important tool to utilize under the appropriate circumstances, and one that every data scientist should have in their tool bag.

References

[1] G. Smith, “Video Game Sales.” *Kaggle*, Oct. 2016, https://www.kaggle.com/datasets/gregorut/videogamesales.

[2] J. Becerra Guerrero, “Sales Cluster & Analysis.” *Kaggle*, Mar. 2021, https://www.kaggle.com/code/jaimebecerraguerrero/sales-cluster-analysis/notebook.