Preprocesamiento de datos de "Video Game Sales with Ratings"

Nicolas Larrañaga Cifuentes[#1], Daniel Augusto Cáceres Salas[*2]

[#]*Facultad de Ingeniería, Universidad Nacional de Colombia*

[1]nlarranagac@unal.edu.co

[2]daacaceressa@unal.edu.co

Bogotá D.C, Colombia

***Abstract*— Este documento provee información y un ejemplo aplicado acerca de la actividad de preprocesamiento de datos en el proceso de KDD. Las técnicas utilizadas en este proyecto son: Muestreo, Normalización, Discretización, Eliminación de ruido, Manejo de valores perdidos y PCA.**

***Keywords*—** KDD, PCA, K-bins, Videojuegos, Ventas.

I. Introducción

Actualmente la industria de los videojuegos se ha apropiado de un amplio porcentaje de del mercado mundial, únicamente en norteamérica el valor del mismo ronda los 18.4 mil millones de dólares. Enmarcado en el contexto de la fuerte competencia generada por las grandes casas diseñadoras de videojuegos, ha surgido una nueva tendencia la cual consiste en apuntar a grupos de nicho basados en la locación del público y el género del videojuego, de esta manera se busca crear una base de clientes frecuentes acostumbrados a adquirir los juegos de un género específico que sean producidos (o en algunos casos distribuidos) por alguna firma en específico.

Este set de datos posee la información de ventas, género, plataforma, calificación (crítica y usuarios) y rating ESRB de videojuegos de los últimos 30 años repartidos en un total de 16719 registros y 16 variables. Debido a la gran cantidad de videojuegos en la actualidad, estos datos se limitan a los juegos desarrollados por compañías reconocidas a nivel mundial, excluyendo los denominados videojuegos independientes o "indie games" los cuales son creados por individuos o pequeños grupos, sin apoyo financiero de distribuidores. La información mencionada es de acceso público y puede ser encontrada en el siguiente enlace: https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings/data

La idea de este proyecto es crear un modelo predictivo que permita a las empresas desarrolladoras de videojuegos obtener un estimado de ventas de un producto nuevo en las diferentes regiones (Japón, Norteamérica, Europa) con base al género, rating y recepción crítica del mismo.

Los conjuntos de datos (original y procesado) junto a los scripts utilizados para el tratamiento pueden ser encontrados en el siguiente repositorio de github: https://github.com/larranaga/UNAL-data-mining.

II. Descripción de las variables

Las variables presentes en el set de datos original son las siguientes:

Variable	Descripción	Tipo	Categoría
Name	Nombre del Juego	String	Nominal
Platform	Consola en la que se juega.	String	Nominal
Year of release	Año de lanzamiento.	Numérico Discreto	Ordinal
Genre	Género (Acción, Deportes, RPG, Shooter, etc)	String	Nominal
Publisher	Firma encargada de distribuir el juego.	String	Nominal
NA sales	Número de ventas en Norteamérica (millones de unidades).	Numérico Discreta	Radio

EU sales	Número de ventas en Europa (millones de unidades).	Numérico Discreta	Radio
JP sales	Número de ventas en Europa (millones de unidades).	Numérica Discreta	Radio
Other Sales	Número de ventas en el resto del mundo, i.e. África, Asia (excluyendo Japón), Australia y Sudamérica (millones de unidades).	Numérica Discreta	Radio
Global Sales	Número de ventas a nivel mundial. (Suma de las anteriores 4 variables).	Numérico Discreta	Radio
Critic Score	Promedio del puntaje dado por los críticos de la página Metacritic.	Numérico Discreta	Radio
Critic Count	Cantidad de críticos que evaluaron y dieron un puntaje al videojuego.	Numérico Discreta	Intervalo
User Score	Promedio del puntaje dado por los suscriptores de la página Metacritic.	Numérico Discreto	Radio
User Count	Cantidad de usuarios que evaluaron y dieron un puntaje al videojuego.	Numérico Discreto	Intervalo
Developer	Firma responsable de la creación del videojuego	String	Nominal
Rating	Clasificación dictada por el ESRB (Early childhood, Everyone, Teen, Mature).	String	Ordinal

Tabla. 1 Descripción de las variables

III. Preprocesamiento

Para el preprocesamiento de este set de datos realizamos los siguientes 8 pasos.

1. Manejo de valores perdidos

El proceso inicia mediante la eliminación de los registros que tengan en la variable 'Name' un valor nulo; en la variable 'User Score' un valor tbd (to be determined); en las variables 'Publisher' y 'Developer' valores nulos. Ya que estos registros contienen demasiada información perdida como para ser utilizados. Una vez finalizada esta eliminación de registros se procede a insertar información en los registros cuyas variables 'Rating' , 'Developer' o 'Publisher y 'Year of release' son nulas. Para 'Developer' y 'Publisher' se inserta el valor no nulo en la otra variable. Para 'Rating' se inserta el valor de la moda del developer en este atributo. Para 'Year of release' se inserta la media de este atributo, esta solución es factible debido a la naturaleza de los datos (los videojuegos llevan un tiempo relativamente corto en haber sido comercializados).

2. Eliminación de ruido

Debido al significado de las variables 'User Score' y 'Critic Score' se pueden encontrar registros sin significado, el valor dado por los usuarios debe ser entre 0 y 10, mientras que el valor dado por la crítica debe estar en 0 y 100. Por lo tanto cualquier registro que contenga un valor fuera de estos rangos en su respectiva variable debe ser eliminado.

3. Normalización

La variable 'Critic Score' es normalizada dividiendo entre 10 su valor para que se encuentre en el mismo rango que la variable 'User Score'. Esto se debe a que ambos atributos están muy relacionados por lo que es conveniente que se encuentren en el mismo rango.

4. Muestreo

El primer cambio sustancial en la información viene al comprimir la cantidad de registros mediante la dimensión 'Platform' al combinar los atributos de los registros del mismo juego en diferentes plataformas, es decir, diferentes registros de un juego se van a convertir en un único registro.

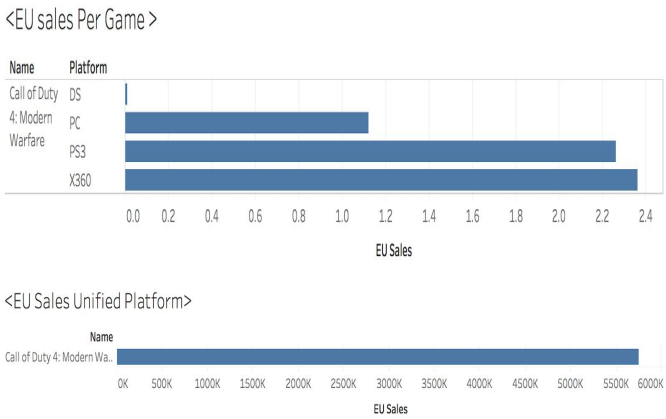

Fig. 1 Combinación de registros por plataforma

Como se puede evidenciar en la figura 1, los 4 registros de diferentes plataformas se convierten en uno solo. Las variables 'Name', 'Year of Release', 'Genre', 'Publisher', 'Developer' y 'Rating' son comunes en todos los registros del mismo juego por lo que en el registro unificado se mantienen iguales.

Las variables 'Sales', 'Critic Count' y 'User Count' se suman ya que este sería el valor total de las ventas por zona del juego.

Por último, las variables 'Critic Score' y 'User Score' se promedian con respecto a la cantidad de registros que tienen valores diferentes a 0.

5. Reducción de dimensionalidad

Al combinar los diferentes registros podemos eliminar el atributo 'Platform' sin ningún problema.

6. Manejo de valores perdidos

Este paso es realizado una vez más ya que aún se encuentran muchos valores nulos en las dimensiones 'User Score' y 'Critic Score'. De los alrededor de 10.000 registros que se tienen hasta el momento, alrededor de la mitad poseen valores nulos en ambas variables. La solución es hallar la media de estos atributos según el developer y asignarla a estos registros. Esto se realiza bajo la suposición que un developer exitoso va a seguir produciendo juegos de la misma calidad, mientras que uno de bajo desempeño se mantendrá en esta categoría.

Aún después de esta gran suposición más de 1.000 datos siguen teniendo valores nulos en las variables 'User Score' y 'Critic Score', esto se debe a que muchos juegos fueron desarrollados por un developer que no posee más juegos registrados en este set de datos.

7. Discretización

Debido a la importancia de valores discretos para algunos algoritmos de asociación y clasificación se optó por realizar una discretización en las variables de tipo 'Sales' (NA, EU, JP y Other). Esta discretización fue realizada con la herramienta Tableau utilizando la técnica de K-BINs con una cantidad diferente de bins recomendada por Tableau para cada una de las variables.

Finalmente la variable 'Global Sales' fue recalculada sumando los nuevos valores de las dimensione de ventas.

8. PCA

Para realizar la reducción de dimensionalidad utilizando el algoritmo PCA se utiliza la librería sklearn de Python. Para el uso de esta librería se debe realizar una conversión de las variables de tipo String a tipo entero (cabe resaltar que esta transformación no cambia el hecho de que estas variables sean de categoría nominal).

Una vez realizado el algoritmo de PCA llamando la función fit() de la librería podemos obtener el arreglo explained_variance_ratio_.cumsum() el cual se puede apreciar en la figura 2.

```
0.9458753748078372
0.9774095185442682
0.9957856275326887
0.9999999265821357
0.99999998549735
0.9999999996067076
0.9999999998755031
0.9999999999934568
0.9999999999978755
0.9999999999994034
0.9999999999996987
0.9999999999999404
0.9999999999999999
0.9999999999999999

Process finished with exit code 0
```

Fig. 2 Valores acumulados del porcentaje de información retenido según la cantidad de dimensiones

Con esta información se concluye que la nueva cantidad de dimensiones en el set de datos va a ser 4, esto se debe a que al ser reducida a esta cantidad de dimensiones aún se conserva el 99.999999266% de la información.

IV. Conclusiones

El conjunto de datos original presentaba problemas con el número de datos faltantes, sin embargo estos valores pudieron ser calculados a costa de perder 6000 datos de 16000. Adicionalmente el análisis de reducción de dimensionalidad permitió ver que las variables tienen una colinealidad alta entre sí, permitiendo la reducción de 16 originales a 4, lo cual da una ventaja para la elaboración de un modelo predictivo como siguiente paso.