

Análisis de información musical mediante técnicas de clustering sobre datos de Spotify

1st Juan Ignacio Vázquez Broquá

Licenciado en Economía
UCA

Buenos Aires, Argentina
juanivazquez@gmail.com

2nd Santiago Guizzardi

Ingeniero Civil
UBA

Buenos Aires, Argentina
santiagoguizzardi@gmail.com

3rd Eric Kalinowski

Ingeniero en Sistemas de Información
UTN

Buenos Aires, Argentina
eric.kalinowski95@gmail.com

En el presente trabajo se realiza un análisis de audio sobre datos provistos por las API de Spotify, con el objetivo entender si las pistas musicales se agrupan naturalmente por géneros a partir de la observación de los distintos atributos de un tema. Para esto se cuenta con tres conjuntos de datos: Audio Tracks, Audio_Features, y Audio_Analysis. Sobre estos se realizaron distintos métodos de clustering comparando como era su comportamiento bajo diferentes condiciones (normalización, estandarizaciones, tratando o sin tratar outliers y sobre distintos métodos de clustering). El mejor resultado fue el obtenido para el dataset consolidado sin transformaciones sobre las variables numéricas y aplicando la transformación de Gower [1] para capturar las distancias de las variables categóricas. Se distinguen géneros con diferencias de bajo nivel marcadas como Death-Metal, Ska y Opera. En contraposición, los géneros menos destacados fueron los asociados a ambient y jazz y singer-songwriter, siendo géneros con gran variedad.

***Index Terms*—clustering, spotify, k-means, PMA, hierarchical clustering, audio, tracks, features, análisis, música**

I. INTRODUCCIÓN

La aparición de grandes repositorios musicales a partir de bases preexistentes en formatos analógico, y la creciente demanda de más y mejores herramientas para clasificar y acceder a contenidos musicales genera una demanda de descriptores eficaces y confiables para la rápida identificación de agrupamientos musicales. Como señala Scaringella [2] a pesar de la popularidad de los "géneros musicales", estos siguen siendo un concepto vago y cambiante, haciendo difícil la clasificación automática de la música. Con este punto de partida, nos proponemos explorar los agrupamientos por género definidos por Spotify, buscando detectar la coincidencia entre las clasificaciones brindadas por la plataforma y los resultados obtenidos a partir de técnicas de clasificación no supervisadas vistas en clase.

II. PREPROCESAMIENTO

Se comenzó realizando un análisis exploratorio de los 3 datasets mencionados, explorando las definiciones de las variables en cada uno, y explorando las distribuciones, presencia

de observaciones atípicas y datos faltantes. Se eliminaron muchas variables irrelevantes asociadas a clasificaciones externas como ser dominios web, número de canción en disco, etc. Se exploró la forma de las distribuciones de las variables e inicialmente se realizaron transformaciones de boxcox, y logarítmicas para corregir sesgos de las distribuciones. Mas adelante estas transformaciones no mejoraron la calidad de las clasificaciones no supervisadas y se siguió trabajando con las variables originales. Asimismo, se probaron dos técnicas de escalamiento de variables: min-max y estandarizado normal. Ninguno de los escalados demostró contribuir a la calidad de las clasificaciones y fueron descartados en los datasets definitivos. Se hizo la prueba de utilizar nuevas variables que resuman los datos, para ellos tomamos el mínimo y el máximo del pitch y del timbre y se hizo la prueba de sumarlos a los datasets de alto nivel y al dataset de bajo nivel, pero no se vio una mejoría notoria, sino que en el caso del dataset de Alto nivel, era un poco peor.

Por último, en el caso de la variable "Key" asociada a la clave musical, siendo esta una variable categórica ordinal de carácter cíclico, se normalizaron las variables en base 2 x π y se crearon las variables adicionales Seno y Coseno para detectar la cercanía de las claves musicales entre sí (por ejemplo, entre Do y Sí, o entre Do y Re).

Para las variables de Audio Análisis se construyeron variables sintéticas del Timbre y Pitch a partir de definiciones compartidas en clase. Como resultado del trabajo preliminar, se decidió trabajar con 4 datasets:

- *Alto Nivel*: Este dataset consolida las variables de alto nivel más valiosas obtenidas de los datasets de Audio Tracks y Audio Features.
- *Bajo Nivel*: Este dataset consolida las variables sintéticas generadas para el dataset de Audio Análisis agrupadas por id de cada canción.
- *Total*: Es el dataset consolidado agrupando la información de Alto y Bajo Nivel.
- *Gower*: Es el dataset obtenido a partir de la transformación de Gower sobre el dataset consolidado. Este dataset es el que más información contiene, por recoger distancias entre variables continuas y categóricas.

III. EXPERIENCIA 0: EXPLORACIÓN JERÁRQUICA

Para los distintos datasets planteados (Tabla I), se evalúan si los datos se agrupan naturalmente por géneros. Para este análisis, primero se evalúa cual es la tendencia al clustering del dataset y como indicador se utiliza el coeficiente de Hopkins.

Luego, se realiza un clustering jerárquico para ver cómo se organizan los datos y si tienen algún agrupamiento. El árbol da una idea de cuan fácil va a ser encontrar clústers. Este gráfico se complementa con la matriz de distancias y el cálculo de coeficiente de correlación cofenética. El cual mide cuanto se parecen las distancias dentro del árbol respecto a las distancias originales.

Por último, se calcula el coeficiente de silhouette promedio de los datos, el silhouette para cada registro y el promedio para cada clúster.

Todo el análisis se realiza para distintas combinaciones de variables, transformaciones de las mismas y distintos métodos y métricas para el clúster jerárquico. Los resultados se presentan en la Tabla II.

Dataset	Descripción
1	Todas las variables numéricas normalizadas
2	Todas las variables numéricas
3	Variables Numericas de tracks y audio_features
4	Variables Numericas de audio_analysis

TABLE I
COMPOSICIÓN DE LOS DISTINTOS DATASETS.

Dataset	Método	Métrica	Hopkins	Coef. Cor. Cofenética	Silhouette Promedio
1	average	euclidean	0.084	0.643	0.033
1	single	euclidean	0.084	0.458	0.033
1	complete	euclidean	0.084	0.500	0.033
2	average	euclidean	0.005	0.880	-0.261
1	average	chebyshev	0.084	0.523	0.033
1	single	chebyshev	0.084	0.348	0.033
1	complete	chebyshev	0.084	0.306	0.033
3	average	euclidean	0.071	0.670	0.006
3	single	euclidean	0.071	0.454	0.006
3	complete	euclidean	0.071	0.623	0.006
4	average	euclidean	0.074	0.670	0.035
4	single	euclidean	0.074	0.613	0.035
4	complete	euclidean	0.074	0.400	0.035

TABLE II
COMPARACIÓN ENTRE DISTINTOS DATASETS, MÉTRICAS Y TRANSFORMACIONES.

Analizando los resultados de cada prueba se puede observar que el dataset dos presenta el coeficiente de Hopkins mas bajo y el coeficiente de correlación cofenética mas alto. Pero, el silhouette promedio resulta negativo.

La combinación que dio los mejores resultados generales fue el dataset uno (los datos numéricos escalados), utilizando el método "promedio" y la métrica "euclidiana". Con esta combinación de variables, transformaciones, métricas y métodos se calcula el silhouette para cada registro y se los agrupar por género. Los resultados se presentan en la Figura 1. Donde los números hacen referencia a los géneros según la siguiente lista: 0: 'ambient', 1: 'classical', 2: 'death-metal', 3:

'drum-and-bass', 4: 'jazz', 5: 'opera', 6: 'singer-songwriter', 7: 'ska', 8: 'trance'.

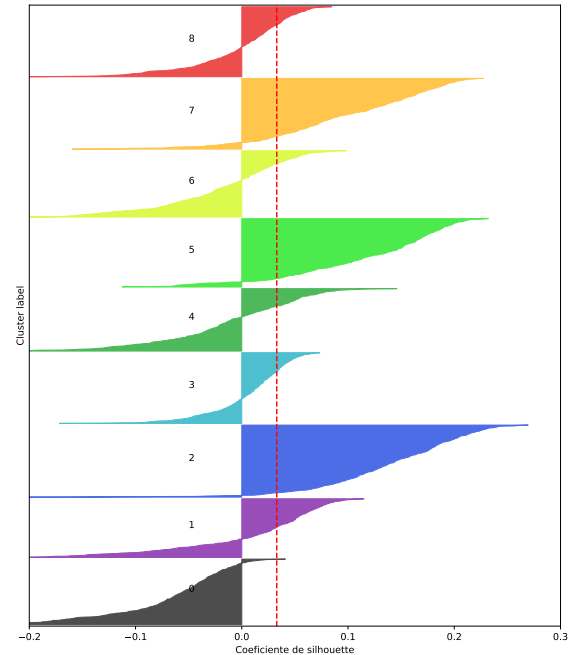


Fig. 1. Silhouette de cada observación agrupados por género.

Analizando el gráfico se puede concluir que existe cierta tendencia al agrupamiento de los datos por género. Aunque se observan grandes diferencias entre los mismo: los géneros que mejor se clusterizan son: 'death-metal', 'opera' y 'ska'. Y el que peor se clusteriza es el 'opera'.

IV. EXPERIENCIA 1: KMEANS

En primer lugar se calculó el estadístico de Hopkins para cada dataset para validar la tendencia natural al clustering en cada caso. Los resultados obtenidos confirman una tendencia natural en todos los datasets (ver tabla III.), siendo el valor de Hopkins 0 el valor óptimo, y 0.5 el escenario inclasificable.

Dataset	Estadístico de Hopkins
Dataset 1	0.0064
Dataset 2	0.2085
Dataset 3	0.0047
Dataset 4	0.011

TABLE III
COMPARACIÓN DEL ESTADÍSTICO DE HOPKINS PARA DISTINTOS DATASETS.

Viendo esta tabla se puede observar que los dataset con mayor tendencia al clustering son el de Alto Nivel y el Total, seguidos por el que se calculo usando Gower.

A. Validación Interna

Para definir a priori la mejor cantidad de clusters en cada caso (k) se calcularon el coeficiente de Silhouette promedio y el mínimo del cuadrado de los residuos (SSE) para cada distintos valores de k. En general, los cuatro datasets arrojaron una caída pronunciada y monótona en ambas curvas. Los valores óptimos elegidos para k fueron los siguientes (ver tabla IV).

Dataset	Óptimo K elegido
Alto Nivel	6
Bajo Nivel	6
Total	9
Gower	7

TABLE IV
VALORES ÓPTIMOS DE CLUSTERS PARA CADA DATASET.

B. Comparación entre Datasets

Como parte de la interpretación de los clusters obtenidos se construyeron matrices de coincidencias entre los clusters óptimos obtenidos en cada dataset. A continuación se comentan algunos casos:

- *Alto Nivel vs Bajo Nivel:* Se ven coincidencias claras entre los datasets, con información de Alto y Bajo Nivel. Esto es esperable, porque gran parte de los features sintéticos del dataset Audio Features son variables obtenidas a partir de los audios.
- *Total vs Gower:* A diferencia de lo esperado, los agrupamientos óptimos obtenidos entre el dataset Total y la transformación de Gower arrojan agrupamientos muy diversos. Esto lo interpretamos como una señal de que las variables categóricas tienen relevancia clasificatoria para los datasets.
- *Gower vs Alto Nivel:* Es interesante observar que el dataset inicial sólo con features de Alto Nivel no tiene puntos de coincidencia destacables con el último dataset (el más elaborado) desarrollado. Esto lo interpretamos como un signo de que la información adicional producida destaca aspectos no recogidos en los features iniciales.

C. Validación Externa

Elaborando ad-hoc un clustering kmeans para los datasets se construyeron matrices de confusión para comparar 9 clusters con los nueve géneros definidos apriori por Spotify. Los resultados indican que existen grupos musicales fácilmente identificables en un grupo como es el caso de Death-Metal y Ska, mientras que otros géneros resultaron pobremente clasificados en todos los casos como es el caso de los géneros ambient y jazz. Los géneros más distintivos tienen rasgos particulares como puede ser la intensidad del audio (loudness) y el tempo musical (time_signature). Por su parte, los géneros peor clasificados son vagos en sus definiciones musicales y cubren amplios espectros como es el caso del jazz.

D. Comentarios Finales

Los resultados obtenidos vía kmeans para los cuatro datasets indican una coincidencia destacable entre los features de alto y bajo nivel, observada en la validación de coincidencias de ambos grupos. Al trabajar con la base más completa (gower, ver figura 2) se identifican fácilmente algunos géneros musicales con aspectos distintivos como son death-metal, Opera, Ska y Trance.

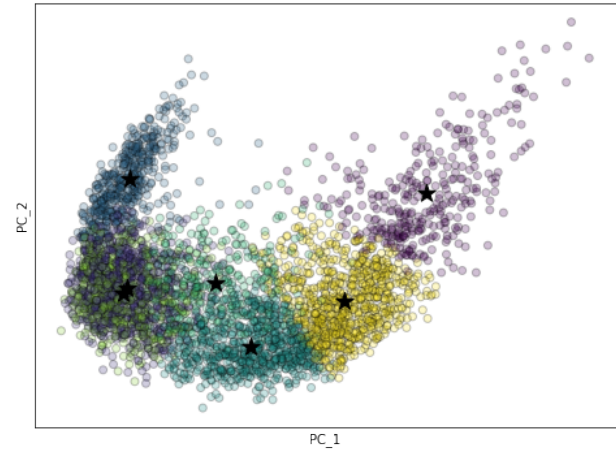


Fig. 2. Clustering usando Gower - Mejor resultado.

En las antípodas, los géneros más difíciles de clasificar resultaron ser ambient, jazz y singer-songwriter, géneros que por definición son amplios y de amplio espectro.

V. OTRAS EXPERIENCIAS

A. PMA

Siguiendo un procedimiento similar se aplicó el algoritmo de KMedoids que por su definición arroja resultados más robustos al escoger como prototipo a una observación del dataset y no a un valor promedio del cluster obtenido. Para este algoritmo se escogió la distancia de Manhattan en contraposición a la distancia euclídea aplicada en la experiencia anterior. Observando los puntos de validación interna, una diferencia notable respecto de los resultados obtenidos con Kmeans es que la curva de Silhouette tuvo en este caso un comportamiento menos predecible al incrementar el valor de k. En todos los casos los valores obtenidos para Silhouette fueron superiores a similares valores de residuos, lo que indica una mejor calidad del algoritmo. Una vez más se exploraron los residuos cuadrados y la evolución del coeficiente de Silhouette para la validación interna y la selección del número óptimo de clusters. La performance de los modelos fue validada externamente aplicando matrices de confusión y comparando el Índice de Rand Ajustado y el Valor Normalizado de van Dongen. En líneas generales los resultados fueron superiores a los obtenidos con el algoritmo anterior.

B. Clustering Jerárquico Aglomerativo

En esta instancia se realizó un clustering jerárquico aglomerativo en los 4 datasets siguiendo el método de linkage.

Se ensayaron distintos metodos y metricas, siendo las de mejor resultado el metodo: 'complete' y la métrica 'euclidean' aplicados sobre el dataset tres: 'todas las variables numéricas normalizadas (Figura 3). La cantidad de clusters se fijó en nueve, la misma cantidad que géneros. Todos los resultados obtenidos fueron peores que los algoritmos utilizados anteriormente (Figura 4).

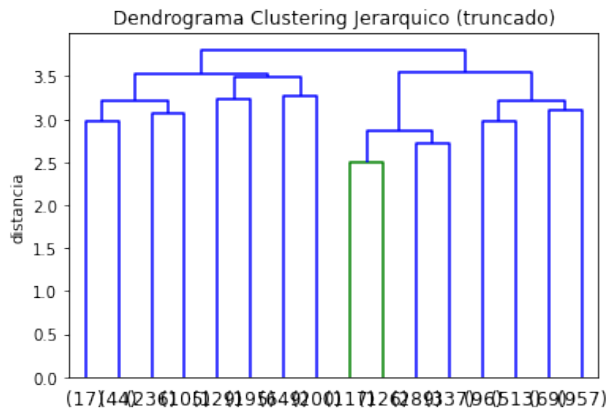


Fig. 3. Árbol Jerárquico.

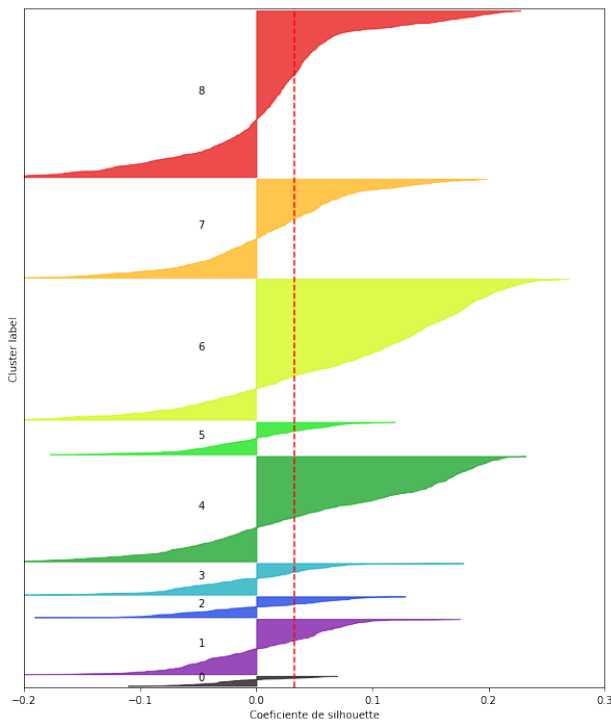


Fig. 4. Coeficiente de Silhouette para el clustering aglomerativo.

VI. CONCLUSIONES

Durante la elaboración del trabajo se exploraron diferentes algoritmos de clasificación no supervisada y se contrastaron los grupos obtenidos con los géneros musicales definidos por Spotify. Se exploraron 3 algoritmos de clasificación, siendo

2 de ellos de clustering prototípico (kmeans y PAM) y uno jerárquico (aglomerativo). El algoritmo que arrojó la mejor performance fue el obtenido con PAM para el dataset Gower. Los géneros musicales mejor clasificados fueron Death-Metal, Opera y en tercer lugar Trance y Ska. Por otro lado, los grupos que peor clasificaron fueron los géneros ambient, jazz y singer-songwriter.

Pensábamos que las diferentes estandarizaciones mejorarían la performance clasificatoria pero esto no ocurrió. Algo que también pudimos observar es que el "esfuerzo" por dotar distintas variables de una distribución normal termino siendo contraproducente, puesto que terminábamos obteniendo resultados mucho peores. De esto ultimo nos llevamos que no siempre el esfuerzo por llevar las variables a una distribución normal se justifica en la práctica.

El mismo resultado lo pudimos ver a partir de las variables que se crearon pensando que podían llegar a aportar una componente discriminante, pero finalmente no se obtuvieron mejoras con las variables que sumamos (principalmente los mínimos y máximos de pitch y del timbre respectivamente), pero estos son pruebas que se realizaron y que desde los mismos resultados negativos aportan al entendimiento del tema.

Finalmente, queremos hacer mención de lo amplio que es el espectro musical y que este trabajo pretende dar una introducción a este análisis y mostrar la variedad de técnicas y herramientas que se tienen para trabajar sobre clusterización, con lo enriquecedoras que pueden ser las conclusiones cuando se explora y se llega a la mejor combinación de valores, y que esta solamente es obtenible a partir de la exploración, las pruebas y el correcto entendimiento de los datos.

REFERENCES

- [1] Gower, J.C.: "Some Distances properties of latent root and Vector methods in Multivariate Analysis", Biometrika, 53, 1966, págs
- [2] Automatic genre classification of music content: a survey - N. Scaringella, G. Zoia, Member, IEEE, D. Mlynec, Member, IEEE