Recomendación de canciones mediante aprendizaje no supervizado

Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Matemática

Santiago Parraguez, Alexandre Poupeau y Yonjairo Sandoval

Resumen

En los años recientes, la música se ha estado vendiendo y consumiendo de forma preferentemente digital. Por esto, la recomendación automática de canciones en las distintas plataformas de reproducción ha sido un problema de gran interés para los desarrolladores. En el presente informe se habla de la importancia de un buen sistema de recomendación de canciones en el mundo moderno, proponiendo un enfoque de selección por *clustering* que toma como base un *data set* de canciones y de gustos por usuario.

1. Introducción

La recomendación de canciones en aplicaciones como Spotify, SoundCloud, iTunes o Youtube a menudo usan métodos de recuperación de información sobre las preferencias del usuario para hacer mejores recomendaciones. Sin embargo, a menudo estas recomendaciones están sometidas a la escucha a gran escala de canciones más populares, dejando de recomendar canciones que tienen un potencial de captar el interés del usuario pero que son menos conocidas.

El objetivo de la recomendación de música es sugerir nuevas canciones a los oyentes considerando sus preferencias. Estas preferencias pueden determinarse de muchas maneras y de distintas fuentes de datos. A partir de los parámetros característicos o representativos de cada canción es posible comparar otras canciones que no hayan sido escuchadas por el usuario y generar la recomendación.

En este estudio se evalúa la opción de recomendar música a partir de los parámetros de las canciones escuchadas por distintos usuarios. Mediante el uso de aprendizaje no supervisado se buscará crear recomendaciones basadas en estos parámetros en vez de utilizar los gustos de otras personas, buscando así realizar una mejor recomendación a la persona.

2. Acercamiento

Million Song Dataset [1] es una colección metadata de un millón de canciones contemporáneas, y aunque no incluye ningún audio, se tienen características representativas de cada canción. Para el desarrollo de este proyecto de investigación se usa un *subset* de 10,000 canciones, ya que resulta poco factible usar un *dataset* tan amplio, aunque el estudio y los métodos utilizados pueden ser extendidos al total de los datos. Así se busca realizar una primera aproximación a este método de recomendación.

Por otra parte, se dispone del *Taste Profile Subset*, un *dataset* complementario que contiene información de reproducción de canciones para distintos usuarios. Este conjunto de datos presentan un *user ID*, *song ID* y un *play count*, con lo que se puede determinar el conjunto de canciones que le gustan a cierto usuario y comparar, dentro de ese mismo conjunto, cuáles son las canciones que prefiere.

El acercamiento que aquí se plantea al problema de recomendación musical es una selección por *clustering* tomando como referencias las canciones más escuchadas por un usuario. El enfoque que se busca es recomendar basándose en las características propias de las canciones, como tiempo, escala o cambios de ritmo. Esto busca evitar caer en círculos de música popular de estilos muy similares basados en lo que escuchan otras personas. Debido a esto es que este método puede resultar un avance significativo a la hora de recomendar música.

3. Metodología

En primer lugar, los *data sets* descritos son sometidos a un pre-procesamiento para describir variables y cantidades importantes en el contexto del problema. Esto ya que en el *data set* de gustos musicales se incluyen canciones que no están en el *subset* de 10,000 canciones, lo que eventualmente puede causar problemas al hacer el *clustering*.

Siendo $\mathcal{D}$ el set de 10,000 canciones, se genera un algoritmo que filtra los datos según un parámetro de entrada y un *threshold* asociado, definido como el número mínimo de canciones que deben existir al mismo tiempo en la data de los gustos musicales del usuario sometido al filtro y $\mathcal{D}$. De esta forma, se genera una métrica variable para filtrar usuarios que tengan una variedad de datos importantes a comparar en $\mathcal{D}$ y se podría evitar al usar el *data set* completo de *Million Song Dataset*. Los resultados de este ejercicio se pueden ver en la tabla 1

Tabla 1: Pre-procesamiento del *dataset* de gustos musicales vs el *subset*

Número mínimo de canciones / usuario	Datos totales	Usuarios	Canciones diferentes
5	151,057	22,840	2,939
8	51,192	5,081	2,166
10	26,600	2,145	1,701
11	19,740	1,459	1,506
12	14,702	1,001	1,351
13	10,994	692	1,189
14	8,342	488	1,062
15	6,550	360	958
16	5,125	265	888
17	4,053	198	811
18	3,288	153	737
19	2,568	113	676
20	2,093	88	635

Para resolver el problema con *clustering*, se realiza un aprendizaje no supervisado en el conjunto de los datos para cada usuario. Se define que cada usuario tiene t canciones de test y el resto de entrenamiento en datos de preferencia de usuarios. Por cada usuario, se realiza un *clustering* usando los *features* de cada canción de entrenamiento escuchada. Luego, para cada centro de los *clusters*, se genera una búsqueda por similitud utilizando las canciones en los 10,000 datos en $\mathcal{D}$ (excepto aquellas que pertenecen a las canciones de entrenamiento del usuario).

Búsqueda por similitud es un método en el que se ordena por distancia sobre $\mathcal{D}$ utilizando las *features*, las N canciones que son mas cercanas al centro de cada *cluster*. Para evaluar ese método, se utilizarán las $K \times N$ canciones mas próximas: si todas pertenecen al conjunto de canciones del test entonces el *average precision* AP_i de usuario i vale 1. Si por el contrario ninguna pertenece al conjunto test, AP_i vale 0. Iterando sobre todos los usuarios se puede obtener un *mean Average Precision* $mAP = \sum_i AP_i$ que será el indicador de la calidad del modelo de recomendación de canciones.

El método recién planteado implica la presencia de varias variables:

- m : el número mínimo de canciones de preferencia por usuario. Ese parámetro define el *dataset* de preferencia utilizado para la evaluación del método.
- Algoritmo de *Clustering*: Se hará uso de $K - Means$.
- K : el número de *clusters*. Tiene que ser fijo durante la evaluación.
- N : el número de canciones mas cercanas tomadas para cada centro de *cluster* para evaluar el modelo.
- t : el número de canciones de test para cada usuario. Se considera la restricción $t < m/2$ para que al menos la mitad de las canciones sean de entrenamiento.

Debido a que el método para evaluar el modelo se basa en las canciones que ya conoce la persona, y a que los datos utilizados no son de las magnitudes ideales, es que los resultados debiesen ser corroborados con usuarios reales, en los cuales se pueda comprobar si es que una canción recomendada es efectivamente de su gusto.

Adicionalmente, en la evaluación del modelo de K - *Means* se usaron $avg\_on\_cluster = False$ y $method =$ "one_if_any" dado que es el que obtiene

mejores resultados considerando que el modelo está limitado por la cantidad de datos comunes entre el *Million Song Subset* y el *Taste Profile Subset*. Esto se corroboró contra los parámetros $avg\_on\_cluster = True$ y $method =$ "$basic$".

4. Resultados Parciales

A partir del código generado se realizaron varias iteraciones cambiando los parámetros mostrados en la sección anterior. Dado que el experimento resultó acotado por la cantidad de datos comunes entre ambos *subsets*, no se dispone de mucha información para evaluar de forma real el comportamiento de la clasificación. Sin embargo, en esta sección se discute la tolerancia del valor de *mean Average Precision (mAP)* al variar los parámetros de entrada del algoritmo.

En la Figura 2 se pueden ver los resultados obtenidos al variar m (número mínimo de datos en ambos *subsets*) y t (cantidad de datos representados por m que fueron usados en el entrenamiento del algoritmo). Como se puede ver se logran alcanzar valores cercanos al 45 %. Se puede ver que los valores son muy sensibles a los parámetros variados, especialmente para distintos t.

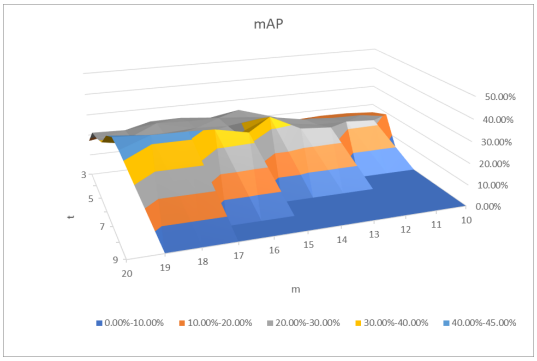

Figura 1: *mean Average Precision* para distintos valores de m y t

Como se puede ver, el valor más alto se obtiene en el mayor m y en el mayor t disponibles en los datos (20 y 9 respectivamente). Para estos valores, resulta interesante evaluar el comportamiento de mAP según el parámetro N, para así poder establecer la flexibilidad del modelo respecto a la cantidad de canciones cercanas tomadas para cada centro de *cluster*. Es lógico pensar que al aumentar las canciones .ªceptadas"dentro de un mismo *cluster*, aumentará también el mAP. El gráfico siguiente permite ver de forma explícita esta sensibilidad.

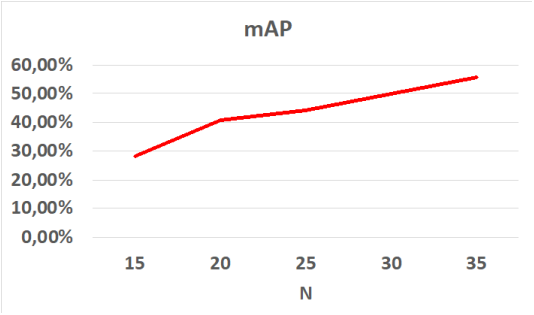

Figura 2: *mean Average Precision* para distintos valores de N (con $m = 20$ y $t = 9$)

5. Conclusiones

Se logra generar una selección de canciones mediante clustering usando el método K - *Means*. Mediante el método utilizado para validar los resultados se obtienen resultados no superiores a un $mAP = 45\,\%$, esto significa que el modelo predijo esa fracción de canciones para un *cluster* de manera efectiva. Este valor se encuentra para $m = 20$ y $t = 9$, lo cual resulta muy específico. En la Figura 2 se puede apreciar que estos valores no se mantienen al variar los parámetros. Esto significa que los resultados obtenidos van a depender profundamente de con cuantas canciones se valide y de el número de canciones que se consideren para hacer la recomendación.

Por otro lado, al variar el número de canciones cercanas a un *cluster* se aprecia un aumento significativo de los valores obtenidos. Esto se explica porque cada centro será más consistente al estar generado por un mayor número de canciones. De esta forma se puede intuir que efectivamente las personas se ven afectadas por parámetros específicos de las canciones que escuchan y no simplemente del estilo que esta canción pueda tener.

Parece evidente que, a mayor número de canciones se utilice para aprender del gusto del usuario, mejores resultados se obtendrán. Mientras que si es que se considera un mayor número de canciones para validar, no resulta tan directo asumir que se logrará encontrar un mayor número de canciones del gusto del

usuario. Esto se explica porque al utilizar muy pocas canciones de validación resulta especialmente difícil recomendar justamente esas canciones, considerando además que ninguna persona tiene un gusto musical totalmente consistente, variando totalmente de una canción a otra. Este problema se aprecia especialmente al utilizar un *dataset* pequeño

Finalmente se considera muy importante realizar un estudio más profundo del caso, utilizando de ser posible el *dataset* completo, y no solo una parte pequeña de el. Con esto se podrían obtener mejores resultados, ya que los obtenidos hasta ahora parecen prometedores. Además es importante corroborar los datos con usuarios reales, y no solo con el método de validación aquí propuesto. Esto arrojaría resultados más precisos respecto a la eficiencia del método a la hora de cumplir con una buena recomendación musical.

El algoritmo generado se puede encontrar en `https://github.com/poupeaua/music-recommendation`.

Referencias

[1] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information. Retrieval: ISMIR 2011*, 2011.