TP1: Data mining en Música

Pablo Riera, Juan E Kamienkowski
Data Mining en Ciencia y Tecnología

6 de octubre de 2020

Music Information Retrieval es el campo que refiere al análisis computacional de señales musicales para aplicaciones tales como el reconocimiento de género o sistemas de recomendación. Muchas de estas aplicaciones se basan en medir una similaridad entre piezas musicales. Utilizaremos los datos de la API de *Spotify* para realizar distintos agrupamientos de piezas musicales con algoritmos de *clustering*.

1. Objetivo

El objetivo general del trabajo es determinar si las pistas musicales (*tracks*) se agrupan naturalmente por géneros a partir de atributos a distintos niveles de observación del tema. Para ello se evaluarán distintos algoritmos de *clustering* y atributos para analizar la relación entre estos atributos y los géneros musicales.

2. Datos y atributos

Se cuenta con tres conjuntos de datos: *tracks*, *audio_features*, y *audio_analysis*. El objetivo será entonces realizar *clustering* para encontrar agrupaciones de los *tracks*. Los *audio_features* tienen descriptores o atributos (*features*) de alto nivel y se pueden usar directamente para el *clustering* previamente seleccionando cuales son útiles. Los datos de *audio_analysis* son descriptores de bajo nivel en formato de una serie temporal multivariada y requieren ser resumidos de alguna forma para obtener un conjunto de descriptores por pista.

3. Preparación de los datos

Los resultados resumidos del pre-informe deben ser incluidos como primer paso en el informe. Brevemente, allí se presentan y exploran los atributos.

4. Generos

Las etiquetas de géneros no siempre se condicen con la información que se puede extraer del audio de la canción, es posible que a veces se consideran otras características musicológicas

o simplemente pueden existir datos mal etiquetados.

a) Evaluar para los distintos datasets (con distintas métricas, normalizaciones, combinaciones, etc) si los datos se agrupan naturalmente por géneros, es decir hacer un análisis utilizando las etiquetas de géneros como si fuesen resultados de un clustering.

b) Identificar los géneros que mejor y peor se clusterizan

5. Clustering

La consigna consiste en aplicar distintos algoritmos de *clustering* y explorar sus hiper-parámetros en base a métricas de validación.

a) Utilizando sólo datos continuos, aplicar *KMeans* sobre ambos conjuntos de datos (*audio_features* y *audio_analysis* con distintas métricas, normalizaciones, combinaciones, etc). Determinar la cantidad de *clusters* utilizando *silhuette* y *SSE*.

b) Evaluar si los agrupamientos son similares para los distintos conjuntos de datos utilizando la matriz de confusión y los índices de *Rand* y *van Dongen* en los casos que correspondan.

c) Evaluar para los mejores casos si los *clusters* se condicen con el género y relacionarlo con la sección anterior.

d) Visualizar los *clusters* y las etiquetas de género en baja dimensión con alguna técnica de reducción (PCA, TSNE, MDS, etc)

e) Discuta brevemente los resultados obtenidos

Repetir los pasos para al menos otros dos algoritmos de *clustering*. En el paso *a)* considerar los hiper-parámetros y la métrica de validación interna que corresponda.

6. Generos Bis

A partir de las experimentos realizados, identificar un conjunto de tres géneros donde los resultados del clustering representan bien a los géneros y un conjunto de tres géneros donde ocurra lo opuesto.

7. Puntos optativos

Elegir al menos uno de los siguiente puntos para completar el informe:

7.1. Nuevos atributos

Considerar el uso de nuevos atributos, tomando las series de tiempo de *audio_analysis* (*timbre* o *pitches*) y computando nuevas variables que resuman los datos más allá de media y desvío. Establecer qué cambios se producen en el *clustering* al incorporar nuevos atributos e interpretar el significado de los mismos.

7.2. Documentar y publicar

Realizar un repositorio que contenga el código y la documentación necesaria para que pueda ser ejecutado por un tercero. Utilizar PMML u otro formato para estandarizar (en el caso de no ser posible reportar las dificultades).

7.3. *Clustering* de secciones dentro de una pista

Los datos de *audio_analysis* permiten ver como se desarrolla un pista en el tiempo. Es posible calcular la matriz de distancias entre los distintos instantes de la pista para obtener la llamada matriz de recurrencia. A partir de ella, es posible generar un *clusteriring* que tome las diferentes partes de una pista, como pueden ser, introducción, estrofa y estribillos por ejemplo.

Para esto, seleccionar una sola pista y realizar *clustering espectral* sobre los datos de *audio_analysis*. La forma de validación más inmediata es escuchar la canción y reconocer las partes que la constituyen.

8. Formato

Les proponemos seguir el formato de publicación en una revista científica, pueden encontrar muchos formatos directamente en *Overleaf*, por ejemplo *IEEE Conference Template for ANCS 2019*. No es obligatorio seguir ese formato, pero si elegir el formato de alguna revista (también pueden encontrarlos disponibles para Word).

Las revistas suelen tener además instrucciones respecto al formato online, desde restricciones en el tamaño de cada sección, en el número de figuras/tablas, las secciones que debe contener, hasta formato de los números, referencias, etc.

Aquí ponemos nuestras restricciones, pero si quieren adaptarlo a alguna revista en particular también vale.

Secciones.

1. Título (máx. 100 caracteres), tiene que ser expresivo (no vale TP1).

2. Resumen (máx. 200 palabras), tiene que contener una descripción de todo el trabajo: Motivación, Antecedentes, Objetivos, Métodos, Resultados y alguna Conclusión.

3. Introducción Comienza con la motivación, sigue con los antecedentes, y termina siempre con un párrafo de objetivos (no es necesario que este dividido en sub-secciones). Típicamente, una vez que motivaron el trabajo y mostraron lo que hay hecho, viene una frase del estilo "Por ende, nos proponemos....º .ªquí nos proponemos...".

4. Métodos Detalle de los métodos a utilizar, en este caso no es necesario profundizar mucho pero pueden enumerarlos y sobre todo es el lugar para incluir cualquier método fuera de lo común que hayan utilizado.

5. Resultados y discusión Aquí se enumeran y discuten los resultados. Es muy importante que no sea una seguidilla de figuras y tablas. Como regla pueden considerar: *"Si una figura no se describe/comenta en el texto es que: O bien está de más y no hace a la historia, o bien se olvidaron de incluirla."*.

6. Conclusiones Comienza generalmente con un resumen muy breve de los principales resultados obtenidos (uniendo distintas secciones), y luego se pasa a conclusiones generales, detallando problemas detectados, posibles explicaciones y trabajo a futuro.

7. Referencias Citas bibliográficas utilizadas durante el reporte. Si son sitios web o repositorios se incluyen generalmente al pie de la página que corresponde y no como cita bibliográfica.

Considerando Introducción, Métodos, Resultados y Conclusiones no deben superar las 5000 palabras. Finalmente, considerando el formato de este trabajo pueden dividirlo en

1. Título
2. Resumen
3. Introducción
4. Experiencia 1: KMeans por ejemplo,
   Métodos
   Resultados y discusión
5. Experiencia 2: Otro algoritmo,
   Métodos
   Resultados y discusión
6. Experiencia N:
7. Conclusiones
8. Referencias

9. Entrega

La fecha límite de entrega es el día Domingo 01/11/2020 a las 23.59hs