Sistemas de Recomendación para Plataformas de Streaming

Jorge Eduardo Enciso Agudelo1, Daniel Figueroa Torres1, Sergio Andrés Rojas Moreno1, Édison Javier Diaz Prieto1, Alexander Luna Ruiz1, Árnol Guerrero Calderón1

1 Sistemas Inteligentes, Maestría en Inteligencia Artificial, Pontificia Universidad Javeriana Bogotá D.C., Colombia

**Resumen.** Este proyecto busca modelar un algoritmo clasificador que permita determinar el género de una canción basado en algunos datos importantes que se extraen de un análisis preliminar de la canción. Existen diferentes variables que, en conjunto, permiten determinar un patrón para identificar el género. Sin embargo, se observa que existen algunos casos donde es imposible encontrar dicho patrón por la amplitud de la definición del género en cuestión. Al final, lo que se quiere es determinar la arquitectura correcta que permita la mayor exactitud en el proceso de sugerir los géneros musicales más importantes.

1. Motivación y Contextualización del Problema

Ha pasado más de una veintena de años desde que en 1999, Shawn Fanning transformara la industria musical al popularizar su servicio Napster. La revolución que introdujo esta aplicación a la forma de comercializar la música e incluso, de hacerla, tiene su impacto en la actualidad, donde debido a la multiplicidad de géneros musicales y a la infinidad de canciones de cada género que son introducidas al mercado diariamente a través de aplicaciones de música como Apple Music, Spotify o Deezer, resulta muy compleja la clasificación de una pieza musical en un género determinado: “según declaraciones de Daniel Ek, CEO de Spotify a medios especializados, se suben 20.000 pistas en Spotify cada 24 horas” (Ighman,2018), por lo que resulta improbable escuchar todas las canciones y lograr realizar una clasificación adecuada.​ Adicionalmente, en la actualidad las plataformas aceptan los covers de las canciones originales, las cuales normalmente hacen parte de otro género musical.

De acuerdo con lo anterior, la revolución digital en la industria musical ha generado la necesidad de crear un nuevo sistema con la capacidad de segmentar de una manera más eficiente los géneros musicales, que actualmente son muy amplios y con características heterogéneas, como en el caso de la música alternativa y la música indie.

* 1. Estado del arte del uso de IA para este o para un problema similar

A partir de la motivación y contextualización del problema descritos en el acápite anterior, se llega a la conclusión que el uso de técnicas de Inteligencia Artificial puede resultar beneficioso en cuanto a los resultados que se obtienen, dada la complejidad y magnitud de los datos disponibles. Se decidió hacer uso de esta técnica con referencia en el artículo “Clasificación de género de música bengalí mediante redes neuronales” el cual tiene como objetivo clasificar los subgéneros que tiene la música bengalí. En este articulo hacen diferentes comparaciones entre diferentes métodos para solucionar este problema de clasificación, en los cuales comparan la regresión lineal, la regresión logística, máquinas de soporte de vectores (SVM), k-NN y la técnica de inteligencia artificial de redes neuronales, donde muestra que el modelo de mayor precisión es el de la red neuronal con un 74% de precisión.

**1.1.1. Inteligencia Artificial (IA)**

Con el fin de introducir el concepto de Inteligencia Artificial, en adelante (IA), esta puede definirse como “la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano” (Rouhiainen, 2018, p.17), o como interpretase Marvin Minsky: “...diseño de máquinas con el objeto de simular el funcionamiento de la mente humana y con capacidad adaptativa y libertad de actuación” (Minsky, 2010).

A partir de esta definición, y dado que “las diferentes técnicas de IA, tales como los algoritmos genéticos, las redes neuronales, la lógica difusa, las hiperheurísticas y en general los sistemas de aprendizaje automático (Machine Learning)” (Azar, Tapia, García y Pérez, 2019, p.1), han obtenido resultados importantes en el procesamiento de Big Data y en la resolución de problemas complejos, se propone solucionar el problema de clasificación de música, objeto del presente artículo, a partir de la implementación de Redes Neuronales Artificiales, teniendo en cuenta su capacidad de generalizar información extraída de datos experimentales (Ponce C., 2010).

**1.1.2. Redes Neuronales Artificiales**

Se conoce como Red Neuronal Artificial (Por sus siglas en inglés, ANN) a un modelo matemático o computacional basado en el aspecto o estructura funcionales de las redes neuronales biológicas. La ANN se compone de neuronas artificiales (nodos) interconectados entre sí.

Una definición adicional de ANN propone que es “una técnica de aprendizaje automático que es generalmente eficaz para extraer características críticas de conjuntos de datos complejos y derivar una función o modelo que expresa esas características” (Pelchat y Gelowitz, 2019, p. 1).

Las ANN normalmente buscan solucionar problemas no lineales, las cuales generalmente proporcionan salidas continuas y realizan funciones simples. Las ANN tienen entradas de datos las cuales son ensambladas al modelo de acuerdo con sus funciones operativas y producen respuestas teniendo en cuenta su función de activación.

El tipo de modelo más simple de neurona fue propuesto por McCulloch y por Pitts (1943) y sigue siendo el modelo más utilizado en las distintas arquitecturas de las redes neuronales. Desde 1943, se han desarrollado modelos mucho más detallados y realistas, tanto para neuronas como para sistemas más grandes en el cerebro, lo que lleva al campo moderno de la neurociencia computacional. Por otro lado, los investigadores en inteligencia artificial y estadística se interesaron en las propiedades más abstractas de las redes neuronales, como su capacidad para realizar cálculos distribuidos, tolerar entradas ruidosas y aprender. Aunque ahora entendemos que otros tipos de sistemas, incluidas las redes bayesianas, tienen estas propiedades, las redes neuronales siguen siendo una de las formas más populares y efectivas de sistema de aprendizaje y son dignas de estudio por derecho propio (Rusell & Norvig, 2010).

Este modelo se compone por múltiples señales de entradas del ambiente exterior, las cuales estas conectadas a una neurona artificial, cada conexión cuenta con un peso el cual busca similar las conexiones sinápticas de una red neuronal biológica. Este peso busca dar la relevancia que tiene cada entrada para la neurona. El nodo básicamente hace una suma pondera del peso multiplicado por las entradas menos el sesgo, el cual lo que busca es darle el umbral adecuado a la neurona. A este resultado se le aplica una función de activación el cual tiene como objetivo limitar la salida de la neurona dentro de un rango de valores razonable. Por último el modelo arroja una señal de salida el cual es el valor final producido por la neurona.

Las redes neuronales están diseñadas a partir de una arquitectura que define como se organiza cada neurona entre sí. Existen distintos tipos de arquitecturas, cada una de estas se debe aplicar dependiendo la problemática que se aborde. Entre estas arquitecturas se pueden encontrar: Arquitectura feedforward de una sola capa, la Arquitectura feedforward multicapa, la Arquitectura recurrente, entre otros.

En este caso nos enfocaremos en la arquitectura feedfoward multicapa. Esta arquitectura se compone por una capa de entrada la cual se encarga de recibir la información, una o varias capas ocultas dependiendo la complejidad del modelo, esta capa es la encargada de extraer patrones asociados con el proceso o sistema que se está analizando y por último una capa de salida la cual se encargar de dar la salida final de la red, el cual es el procesamiento realizado por los nodos de las anteriores capas. Este tipo de arquitectura normalmente es utilizado para: clasificación de patrones, identificación de sistemas, control de procesos, optimización, robótica, entre otros.

Las ANN se caracterizan porque tienen la capacidad de aprender, el cual consiste en aplicar una serie de pasos ordenados para ajustar el peso y el sesgo de cada neurona, con el propósito de generalizar la solución producida por sus salidas. Existen distintos tipos de aprendizaje entre los cuales esta: el Aprendizaje supervisado, aprendizaje no supervisado, el aprendizaje por refuerzo, entre otros.

Para abordar esta problemática el modelo se acogerá a utilizar el aprendizaje supervisado. Este consiste en aprender de los datos que se le proporcionan al modelo para entrenar. Estos datos deben estar categorizados para que el modelo al momento de entrenar pueda comparar su salida con el resultado esperado y así por medio de la función de perdida y la función de optimización, el modelo pueda ajustar sus parámetros para disminuir el error que existe entre el dato que se predijo con el dato que se esperaba.

1. Descripción de la tarea

Se quiere entrenar una ANN que tenga la capacidad de clasificar 10 diferentes géneros con la mejor exactitud posible, dichos géneros serán las clases para evaluar dado el dataset seleccionado de la plataforma de Kaggle. Este dataset fue utilizado en un hackathon y el ganador obtuvo un resultado de aproximadamente 0.8 de exactitud, por lo que se presentan retos importantes en la adecuación, balanceo y ajuste de los datos.

* 1. Visión General

Basado en que actualmente existen muchas canciones en el mundo, existen diferentes características que pueden describir el género. Entre ellas tenemos 'Popularidad', 'bailabilidad', 'energía', 'clave', 'sonoridad', 'acústica', 'instrumentalidad', 'viveza', 'valencia', 'tempo', entre otros. Lo que se puede notar de estas características es que, a pesar de estar relacionadas con la música, no necesariamente se conectan a la hora de describir un género musical porque en la lógica humana, por ejemplo, la popularidad no necesariamente puede definir un género, o la bailabilidad, pues, claramente muchos de estos géneros musicales pueden compartir estas características. En este punto es donde se hace necesario el reto de seleccionar correctamente la información para poder llegar a una clasificación efectiva.

Por otro lado, también vale la pena recalcar que la motivación para realizar este proyecto radica en la importancia de reconocer los estilos propios de los diferentes géneros musicales en los diferentes países. Colombia tiene una cultura musical amplia, diferentes géneros que pueden ser muy genéricos, por ejemplo, la cumbia, pero si comparamos las características que tiene este género en Colombia con el estilo que se puede presentar en Argentina o en México, se presentaran diferencias importantes, lo mismo con un género como la Salsa, pues a pesar de que Colombia tiene varios grupos musicales que manejan el género, no es posible comparar los estilos con países como Cuba, Puerto Rico o Republica Dominicana.

Finalmente, del dataset seleccionado hemos analizado asuntos de similitud en medidas, denotando que un entrenamiento de la red con variables que se comportan de la misma forma no resulta favorable para una correcta clasificación, también se encontró un desbalanceo en las diferentes clases, por lo que se tiene que tomar una decisión importante sobre el número de muestras que hay para cada uno de los géneros o clases que se quieren clasificar, con el objeto de lograr resultados coherentes y favorables,

* 1. Restricciones

Tras la visión general presentada, se pueden notar varias restricciones presentes desde el simple hecho de la selección de los datos de valor del dataset. Inicialmente, se requiere hacer un balanceo de los datos, pues de las diferentes clases se ven unos porcentajes desproporcionados que producen fallas a la hora de querer entrenar la red. Adicionalmente, como parte de un análisis de los diferentes géneros, para los casos de la música alternativa y lo que encierra el género indie, hemos encontrado que dentro de ese mismo género siempre van a existir subgéneros muy marcados, por lo tanto, si queremos clasificar un género tan variante, va a ser muy complicado lograr una exactitud del modelo que sea significativa dado que existen otros géneros que presentan características mucho más evidentes. Además, tenemos el caso del género Folk y Country, en donde, técnicamente la música folk es la que emplea instrumentos y formas tradicionales de un lugar para crear nuevos temas musicales, así que puede haber música folk de cualquier origen, en tanto reconozca ese origen en los ritmos y los instrumentos que se emplean. Y la música country es la música folk rural de Estados Unidos, especialmente del centro y oeste del país, es decir, un subgénero. De lo anterior, se pudo definir que el dataset presenta algunas incongruencias con respecto a la selección de los géneros que se querían clasificar, esa situación puede entenderse, ya que hacía parte del reto del hackathon en donde se empleó.

* 1. Alcance

Dada la limitación de tener un sistema que sea capaz de clasificar géneros tan genéricos como el Folk, Alternativo o Indie, es importante tomar decisiones con respecto a los datos del dataset que realmente nos permitan tener exactitud en el modelo planteado. Se realizan varias pruebas con modificaciones leves en la escogencia de los atributos a tener en cuenta en el entrenamiento y posterior clasificación. Con un conocimiento en el alcance que hubo a nivel del concurso donde originalmente se usó este dataset, se pretende lograr mejores resultados que la media en la exactitud del modelo final, sin embargo, se esperan muchos mejores resultados con otros métodos que no sean redes neuronales artificiales, por ejemplo, los algoritmos genéticos.

1. Análisis de potencialidades para el uso de herramientas de Inteligencia Artificial
2. Análisis de viabilidad de validación experimental

A continuación, se describen los métodos empleados para el tratamiento de los datos, preprocesamiento, selección de características, aplicación de Redes Neuronales Artificiales y validación en el proceso de clasificación de los géneros musicales de acuerdo con las características escogidas.

* 1. Selección de la base de datos.

Para la selección de la base de datos, se tomó un dataset de un concurso de la página de machinehack.com en asociación de la plataforma de concursos Kaggle. las cual tiene como objetivo hacer concursos para la comunidad de científicos de datos para que puedan aplicar los conocimientos relacionados a la disciplina de ciencia de datos.

El dataset es un extracto de las canciones más famosas en el año 2020 que se tomó como referencia para el concurso. El dataset de entrenamiento consiste en 17,000 registros que contienen atributos de las canciones. El dataset está compuesto por 14 variables numéricas que describen características específicas de la canción como la instrumentalidad, el tempo, entre otros, y fueron extraídas a priori haciendo uso de software sofisticado de edición de música como Premier Studio. Por otro lado, también se incluyeron variables como popularidad de la canción la cual fue obtenida directamente de la plataforma que suministró la base de datos. Además de estas variables cuantitativas se incluyeron 2 variables cualitativas: nombre de la canción y autor, más la variable de respuesta que es el género de las canciones.

El notebook con el preprocesamiento de datos y los modelos de redes neuronales se encuentran localizados en este link de github : *https://github.com/Jorgeduardo13/SistemasInteligentes*

* 1. Preprocesamiento de los datos

Para la limpieza de los datos se realizó una exploración de datos, en donde tenía el objetivo de encontrar datos atípicos o revisar los datos faltantes de los datos. En primera instancia, se identificó alrededor del 5% de datos faltantes en la variable del ranking de popularidad y en la variable de instrumentalidad.

Por otro lado, también se consideró generar un balance de las clases debido a que existe una cantidad desproporcional entre la clase mayoritaria en comparación a la minoritaria.

**4.2.1. Imputación de valores**

Debido a que el porcentaje de faltantes no era significativamente alto, se decidió en imputar los valores. En este caso se usó un Random Forest con entrada las otras variables para poder predecir los registros faltantes de las dos variables.

**4.2.2. Estandarización de las variables**

Con el fin de poder usar los algoritmos de Machine Learning y de Deep Learning se decidió en estandarizar todas las variables cuantitativas usando la inversa de una función uniforme, o como bien se conoce, usando el algoritmo de min max.

* 1. Exploración de variables

Para esta sección se un análisis por variable y entre variables para poder encontrar datos atípicos y relaciones entre variables que puedan mejorar el desempeño de los modelos. En la tabla de correlaciones se puede apreciar que la mayoría de las variables no tiene una correlación lineal. A excepción de la variable de energía y de ruido(loudness) que tienen un coeficiente de Pearson de 0.77.

Gráfico, Histograma

Descripción generada automáticamente

Tabla 1. Resultados iniciales naive Bayes

* 1. Generación de modelo base

Calendario

Descripción generada automáticamente

Pantalla de computadora con números

Descripción generada automáticamente con confianza media

Tabla 3. Resultados iniciales Random Forrest

Tabla 2. Resultados iniciales naive Bayes

Debido a que este problema es de multi clase y contiene 10 posibles etiquetas, un F1 score de 0.5 es muy optimista ya que no se está tratando con un modelo binario. Por esta razón, se propuso tomar como línea base un modelo de predicción simple o ingenuo que no tenga gran cantidad de híper parámetros para ajustar. En este caso se optó por un modelo de ingenuo bayes, el cual está basado en probabilidad condicional. Además, se incluyó un modelo más sofisticado para que se tenga una idea de cómo es el comportamiento de los datos bajo otro modelo.

En las tablas 2 y 3 se aprecia que los resultados de los modelos iniciales o base tienen un F1 Score de 0.31 y 0.47, lo cual quiere decir que los modelos que se tienen que elaborar con las técnicas de aprendizaje profundo y más adelante con algoritmos genéticos tienen que mejorar estos resultados.

* 1. Implementación de Redes Neuronales

Después de tener una línea de base para poder comparar los desempeños del modelo con algoritmos de aprendizaje profundo, se procede a crear la topología de la red neuronal. Para la implementación de la red neuronal se implementó la librería de Python llamada tensorflow.keras con la cual se puede diseñar la topología de una red neuronal fácilmente.

* 1. Protocolo experimental

La validación experimental se llevará utilizando diferentes características de la base de datos, donde se varían las características los insumos de entrada de la red neuronal en función de maximizar la precisión de la clasificación y poder concluir cuales son las características que le generan mayor valor al modelo.

Para cada experimento se realizó una búsqueda de grilla para encontrar los mejores híper parámetros para la topología de la red neuronal. Cada experimento de los mencionados en la tabla a continuación tenía el mismo campo de búsqueda por lo cual se evaluaban todos los escenarios.

En cada búsqueda de grilla las variables a cambiar eran: Números de neuronas de la primera capa, tasa de dropout en una capa intermedia, función de activación de las capas internas, épocas de entrenamiento y conjunto(batch) con la que se entrenaba la base de datos.

|  |  |  |
| --- | --- | --- |
| Escenario | Características | Best F1 Score |
| 1 | Linea Base Random Forrest. | 0.47 |
| 2 | Red neuronal predeterminada. | 0.43 |
| 3 | Balanceando la base de datos. | 0.51 |
| 4 | Balanceando la base de datos, incluyendo las variables de texto vectorizadas. | 0.57 |
| 5 | Balanceando la base de datos, usando técnicas de reducción de dimensionalidad de componentes principales. | 0.41 |
| 6 | Balanceando la base de datos, usando técnicas de reducción de dimensionalidad de componentes principales, incluyendo las variables de texto vectorizadas. | 0.43 |

*Tabla 4. Resumen de los resultados de la red neuronal.*

Cada ejecución se utilizó un conjunto de entrenamiento y de prueba con los 17,000 los cuales estaban distribuidos 70% y 30% respectivamente. Por cada escenario se corrió una búsqueda de grilla con todo el espacio de búsqueda mencionado anteriormente. Para los algoritmos con mejores resultados se calculó la matriz de confusión para compararlo con los modelos de línea base.

Imagen de la pantalla de un celular con letras y números

Descripción generada automáticamente con confianza baja

Tabla 5. Resultados mejor red neuronal

1. **Conclusiones**

La calidad de los datos de entrada es el paso más importante de un modelo de machine learning. No importa si el modelo es una red neuronal o una simple regresión logística. Si la información que entra al modelo tiene una alta calidad, el modelo obtendrá resultados de calidad. Esto se puede apreciar ya que al incluir las variables de texto con un vectorizado y al haber balanceado las clases, el desempeño de los modelos mejoro independientemente el modelo que fue ejecutado.

Al ser un problema de multi clase, el desempeño de todas las clases no va a ser uniforme, siempre habrá clases que se predicen mejor que otras y este caso no es la excepción. Por ejemplo, en este caso paso con la clase indie. Esta categoría musical al tener tanta variedad es su tipo de canciones, dificulta que se pueda diferenciar de otras clases si se ponen juntas. Esto se puede apreciar ya que el máximo f1 score que fue obtenido fue con el algoritmo de Random Forrest con un 0.07.

Referencias

1. (Cambiar)

2. Anonymous "Development of Neural Networks for Noise Reduction," International Arab Journal of Information Technology, vol. 7, (3), pp. 289-294, 2010. Available: <http://search.ebscohost.com.ezproxy.javeriana.edu.co:2048/login.aspx?direct=true&db=edswsc&AN=000280022900009&lang=es&site=eds-live>.

3. Azar, M; Tapia, M; García, J; Pérez,A. Inteligencia Artificial de las Cosas. (2019).

4. (Cambiar) van Leeuwen, J. (ed.): Computer Science Today. Recent Trends and Developments. Lecture Notes in Computer Science, Vol. 1000. Springer-Verlag, Berlin Heidelberg New York (1995)

5. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996).

6. Pelchat, N; Gelowitz, Craig. Neural Network Music Genre Classification. IEEE Canadian Conference of Electrical and Computer Engineering (CCECE) (2019).

7. Rouhiainen, L. Inteligencia Artificial. 101 cosas que debes saber hoy sobre nuestro futuro. Editorial Planeta (2018)

8. Minsky, M. La Máquina de las Emociones. Barcelona: Debate. (2010).

9. Ponce C., P. Inteligencia Artificial con Aplicaciones a la Ingeniería. México: Algaomega. (2010).

10. Rusell, S., & Norvig, P. Artificial Intelligence A Modern Approach. New Jersey: Prentice Hall. (2010).

11. M. A. Al Mamun, I. Kadir, A. S. A. Rabby and A. Al Azmi, "Bangla Music Genre Classification Using Neural Network," 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), 2019, pp. 397-403, doi: 10.1109/SMART46866.2019.9117400.